Learning to Predict Arbitrary Quantum Processes

Hsin-Yuan Huang,[1,*] Sitan Chen,[2] and John Preskill[1,3]

[1] *Institute for Quantum Information and Matter and Department of Computing and Mathematical Sciences, Caltech, Pasadena, California, USA*

[2] *Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, California, USA*

[3] *AWS Center for Quantum Computing, Pasadena, California, USA*

(Received 14 April 2023; revised 11 November 2023; accepted 15 November 2023; published 6 December 2023)

We present an efficient machine-learning (ML) algorithm for predicting any unknown quantum process $\mathcal{E}$ over n qubits. For a wide range of distributions $\mathcal{D}$ on arbitrary n-qubit states, we show that this ML algorithm can learn to predict any local property of the output from the unknown process $\mathcal{E}$, with a small average error over input states drawn from $\mathcal{D}$. The ML algorithm is computationally efficient even when the unknown process is a quantum circuit with exponentially many gates. Our algorithm combines efficient procedures for learning properties of an unknown state and for learning a low-degree approximation to an unknown observable. The analysis hinges on proving new norm inequalities, including a quantum analogue of the classical Bohnenblust-Hille inequality, which we derive by giving an improved algorithm for optimizing local Hamiltonians. Numerical experiments on predicting quantum dynamics with evolution time up to 10^6 and system size up to 50 qubits corroborate our proof. Overall, our results highlight the potential for ML models to predict the output of complex quantum dynamics much faster than the time needed to run the process itself.

DOI: 10.1103/PRXQuantum.4.040337

I. INTRODUCTION

Learning complex quantum dynamics is a fundamental problem at the intersection of machine learning (ML) and quantum physics. Given an unknown n-qubit completely positive trace-preserving (CPTP) map $\mathcal{E}$ that represents a physical process happening in nature or in a laboratory, we consider the task of learning to predict functions of the form

$$f(\rho, O) = \mathrm{tr}(O\mathcal{E}(\rho)), \tag{1}$$

where ρ is an n-qubit state and O is an n-qubit observable. Related problems arise in many fields of research, including quantum machine learning [1–10], variational quantum algorithms [11–17], machine learning for quantum physics [18–29], and quantum benchmarking [30–36]. As an example, for predicting outcomes of quantum experiments [8,37,38], we consider ρ to be parameterized by a classical input x, $\mathcal{E}$ is an unknown process happening in the lab, and O is an observable measured at the end of the experiment. Another example is when we want to use a quantum ML algorithm to learn a model of a complex quantum evolution with the hope that the learned model can be faster [7,11,12].

As an n-qubit CPTP map $\mathcal{E}$ consists of exponentially many parameters, prior works, including those based on covering number bounds [4,7,8,37], classical shadow tomography [33,39], or quantum process tomography [30–32], require an exponential number of data samples to guarantee a small constant error for predicting outcomes of an arbitrary evolution $\mathcal{E}$ under a general input state ρ. To improve upon this, recent works [4,7,8,37,40] have considered quantum processes $\mathcal{E}$ that can be generated in polynomial time and shown that a polynomial amount of data samples suffices to learn $\mathrm{tr}(O\mathcal{E}(\rho))$ in this restricted class. However, these results still require exponential computation time.

In this work, we present a computationally efficient ML algorithm that can learn a model of an arbitrary unknown n-qubit process $\mathcal{E}$, such that, given ρ sampled from a wide range of distributions over arbitrary n-qubit states and any O in a large physically relevant class of observables, the ML algorithm can accurately predict $f(\rho, O) = \mathrm{tr}(O\mathcal{E}(\rho))$. See Fig. 1 for an illustration. The ML model can predict outcomes for highly entangled states ρ after learning from a training set that only contains data for

[*] hsinyuan@caltech.edu

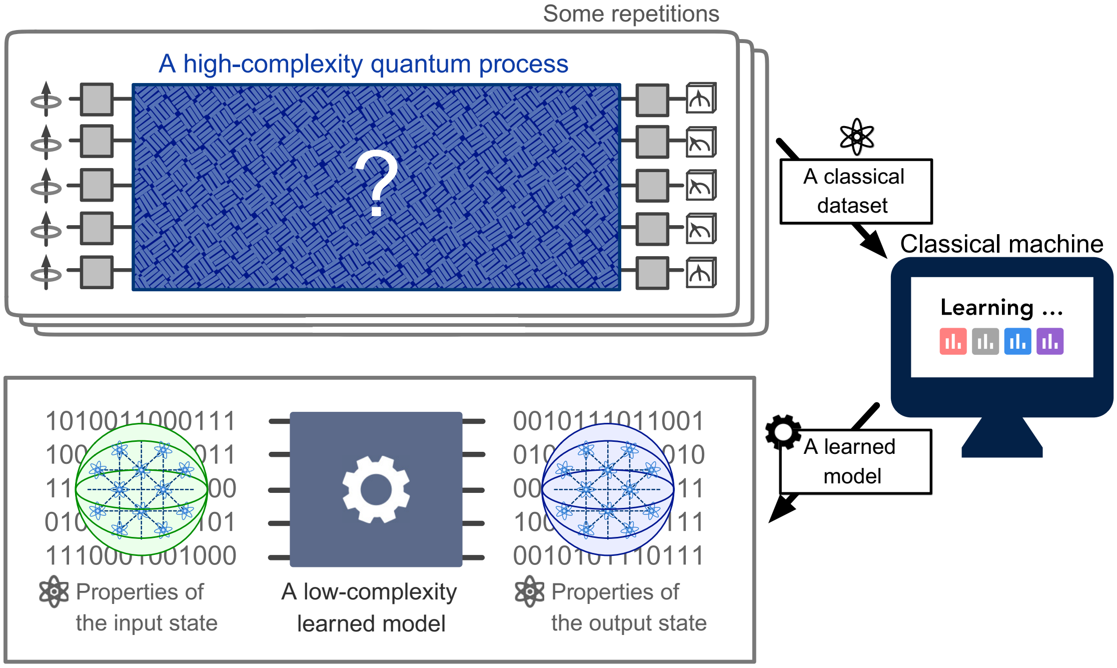

FIG. 1. Learning to predict an arbitrary unknown quantum process $\mathcal{E}$. Consider an unknown quantum process $\mathcal{E}$ with arbitrarily high complexity, and a classical dataset obtained from evolving random product states under $\mathcal{E}$ and performing randomized Pauli measurements on the output states. We give an algorithm that can learn a low-complexity model for predicting the local properties of the output states given the local properties of the input states.

random product input states and randomized Pauli measurements on the corresponding output states. The training and prediction of the proposed ML model are both efficient even if the unknown process $\mathcal{E}$ is a Hamiltonian evolution over an exponentially long time, a quantum circuit with exponentially many gates, or a quantum process arising from contact with an infinitely large environment for an arbitrarily long time. Furthermore, given few-body reduced density matrices of the input state ρ, the ML algorithm uses only classical computation to predict output properties $\text{tr}(O\mathcal{E}(\rho))$.

The proposed ML model is a combination of efficient ML algorithms for two learning problems: (1) predicting $\text{tr}(O\rho)$ given a known observable O and an unknown state ρ, and (2) predicting $\text{tr}(O\rho)$ given an unknown observable O and a known state ρ. We give sample-efficient and computationally efficient learning algorithms for both problems. Then we show how to combine the two learning algorithms to address the problem of learning to predict $\text{tr}(O\mathcal{E}(\rho))$ for an arbitrary unknown n-qubit quantum process $\mathcal{E}$. Together, the sample and computational efficiency of the two learning algorithms implies the efficiency of the combined ML algorithm.

In order to establish the rigorous guarantee for the proposed ML algorithms, we consider a different task: optimizing a k-local Hamiltonian $H = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \alpha_P P$. We present an improved approximate optimization algorithm that finds either a maximizing or minimizing state $|\psi\rangle$ with a rigorous lower or upper bound guarantee on the energy $\langle\psi|H|\psi\rangle$ in terms of the Pauli coefficients α_P of H. The rigorous bounds improve upon existing results on optimizing k-local Hamiltonians [41–44]. We then use the improved optimization algorithm to give a constructive proof of several useful norm inequalities relating the spectral norm $\|O\|$ of an observable O and the ℓ_p norm of the Pauli coefficients α_P associated with the observable O. The proof resolves a recent conjecture in Ref. [45] about the existence of quantum Bohnenblust-Hille inequalities. These norm inequalities are then used to establish the efficiency of the proposed ML algorithms.

II. LEARNING QUANTUM STATES, OBSERVABLES, AND PROCESSES

Before proceeding to state our main results in greater detail, we informally describe the learning tasks discussed in this paper: what do we mean by learning a quantum state, observable, and process?

A. Learning an unknown state

It is possible, in principle, to provide a complete classical description of an n-qubit quantum state ρ. However, this would require an exponential number of experiments, which is not practical at all. Therefore, we set a more modest goal: to learn enough about ρ to predict many of its physically relevant properties. We specify a family of target observables $\{O_i\}$ and a small target accuracy ϵ. The learning procedure is judged to be successful if we can predict the expectation value $\mathrm{tr}(O_i\rho)$ of every observable in the family with ϵ error.

Suppose that ρ is an arbitrary and unknown n-qubit quantum state, and that we have access to N identical copies of ρ. We acquire information about ρ by measuring these copies. In principle, we could consider performing collective measurements across many copies at once. Or we might perform single-copy measurements sequentially and *adaptively*; that is, the choice of measurement performed on copy j could depend on the outcomes obtained in measurements on copies $1, 2, 3, \ldots, j-1$. The target observables we consider are *bounded-degree observables*. A bounded-degree n-qubit observable O is a sum of local observables (each with support on a constant number of qubits independent of n) such that only a constant number (independent of n) of terms in the sum act on each qubit. Most thermodynamic quantities that arise in quantum many-body physics can be written as a bounded-degree observable O, such as local observables, few-body correlation functions, geometrically local Hamiltonians, and the average magnetization.

In the learning protocols discussed in this paper, the measurements are neither collective nor adaptive. Instead, we fix an ensemble of possible single-copy measurements, and for each copy of ρ, we independently sample from this ensemble and perform the selected measurement on that copy. Thus, there are two sources of randomness in the protocol—the randomly chosen measurement on each copy and the intrinsic randomness of the quantum measurement outcomes. If we are unlucky, the chosen measurements and/or the measurement outcomes might not be sufficiently informative to allow accurate predictions. We settle for a protocol that achieves the desired prediction task with a high success probability.

For the protocol to be practical, it is highly advantageous for the sampled measurements to be easy to perform in the laboratory, and easy to describe in classical language. The measurements we consider, *random Pauli measurements*, meet both of these criteria. For each copy of ρ and each of the n qubits, we choose uniformly at random to measure one of the three single-qubit Pauli observables X, Y, or Z. This learning method, called *classical shadow tomography*, was analyzed in Ref. [46], where an upper bound on the sample complexity (the number N of copies of ρ needed to achieve the task) was expressed in terms of a quantity called the *shadow norm* of the target observables.

In this work, using a new norm inequality derived here, we improve on the result in Ref. [46] by obtaining a tighter upper bound on the shadow norm for bounded-degree observables. The upshot is that, for a fixed target accuracy ϵ, we can predict all bounded-degree observables with spectral norm less than B by performing random Pauli measurement on

$$N = \mathcal{O}(\log(n)B^2/\epsilon^2) \tag{2}$$

copies of ρ. This result improves upon the previously known bound of $\mathcal{O}(n\log(n)B^2/\epsilon^2)$. Furthermore, we derive a matching lower bound on the number of copies required for this task, which applies even if collective measurements across many copies are allowed.

B. Learning an unknown observable

Now suppose that O is an arbitrary and unknown n-qubit observable. We also consider a distribution $\mathcal{D}$ on n-qubit quantum states. This distribution, too, need not be known, and it may include highly entangled states. Our goal is to find a function $h(\rho)$ that predicts the expectation value $\mathrm{tr}(O\rho)$ of observable O on state ρ with a small mean squared error:

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho) - \mathrm{tr}(O\rho)|^2 \leq \epsilon.$$

To define this learning task, it is convenient to assume that we can access training data of the form

$$\{\rho_\ell, \mathrm{tr}(O\rho_\ell)\}_{\ell=1}^{N}, \tag{3}$$

where ρ_ℓ is sampled from distribution $\mathcal{D}$. In practice, though, we cannot directly access the exact value of the expectation value $\mathrm{tr}(O\rho_\ell)$; instead, we might measure O multiple times in state ρ_ℓ to obtain an accurate estimate of the expectation value. Furthermore, we do not necessarily need to sample states from $\mathcal{D}$ to achieve the task. We might prefer to learn about O by accessing its expectation value in states drawn from a different ensemble.

A crucial idea of this work is that we can learn O efficiently if distribution $\mathcal{D}$ has suitably nice features. Specifically, we consider distributions that are invariant under single-qubit Clifford gates applied to any one of the n qubits. We say that such distributions are *locally flat*, meaning that the probability weight assigned to an n-qubit state is unmodified (i.e., the distribution appears flat) when we locally rotate any one of the qubits. Locally flat distributions include random product states, ground and thermal states of random local Hamiltonians, and any state that is generated by a circuit, where the last circuit layer consists of random single-qubit gates. Furthermore, any distribution that is at most polynomially far from a locally flat distribution (measured in terms of the maximum likelihood ratio) can be predicted efficiently and accurately.

An arbitrary observable O can be expanded in terms of the Pauli operator basis:

$$O = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \alpha_P P. \tag{4}$$

Though there are 4^n Pauli operators, if distribution $\mathcal{D}$ is locally flat and O has a constant spectral norm, we can approximate the sum over P by a truncated sum,

$$O^{(k)} = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}:\ |P| \leq k} \alpha_P P, \tag{5}$$

including only the Pauli operators P with weight $|P|$ up to k, those acting nontrivially on no more than k qubits. The mean squared error incurred by this truncation decays exponentially with k. Therefore, to learn O with mean squared error ϵ, it suffices to learn this truncated approximation to O, where $k = \mathcal{O}(\log(1/\epsilon))$. Furthermore, using norm inequalities derived in this paper, we show that, for the purpose of predicting the expectation value of this truncated operator, it suffices to learn only a few relatively large coefficients α_P, while setting the rest to zero. The upshot is that, for a fixed target error ϵ, an observable with constant spectral norm can be learned from training data with size $\mathcal{O}(\log n)$, where the classical computational cost of training and predicting is $n^{\mathcal{O}(k)}$.

Usually, in machine learning, after learning from a training set sampled from a distribution $\mathcal{D}$, we can only predict new instances sampled from the same distribution $\mathcal{D}$. We find, though, that, for the purpose of learning an unknown observable, there is a particular locally flat distribution $\mathcal{D}'$ such that learning to predict under $\mathcal{D}'$ suffices for predicting under any other locally flat distribution as well as any other distribution that is at most polynomially far away from a locally flat distribution. Namely, we sample from the n-qubit state distribution $\mathcal{D}'$ by preparing each one of the n qubits in one of the six Pauli operator eigenstates $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\}$, chosen uniformly at random. Pleasingly, preparing samples from $\mathcal{D}'$ is not only sufficient for our task, but also easy to do with existing quantum devices.

After training is completed, to predict $\mathrm{tr}(O\rho)$ for a new state ρ drawn from distribution $\mathcal{D}$, we need to know some information about ρ. State ρ, like operator O, can be expanded in terms of Pauli operators, and when we replace O by its weight-k truncation, only the truncated part of ρ contributes to its expectation value. Thus, if the k-body reduced density matrices (RDMs) for states drawn from $\mathcal{D}$ are known classically then the predictions can be computed classically. If the states drawn from $\mathcal{D}$ are presented as unknown quantum states then we can learn these k-body RDMs efficiently (for small k) using classical shadow tomography and then proceed with the classical computation to obtain a predicted value of $\mathrm{tr}(O\rho)$.

C. Learning an unknown process

Now suppose that $\mathcal{E}$ is an arbitrary and unknown quantum process mapping n qubits to n qubits. Let $\{O_i\}$ be a family of target observables and $\mathcal{D}$ be a distribution on quantum states. We assume the ability to repeatedly access $\mathcal{E}$ for a total of N times. Each time, we can apply $\mathcal{E}$ to an input state of our choice, and perform the measurement of our choice on the resulting output. In principle, we could allow input states that are entangled across the N channel uses, and allow collective measurements across the N channel outputs. But here we confine our attention to the case where the N inputs are unentangled, and the channel outputs are measured individually. Our goal is to find a function $h(\rho, O)$ that predicts, with a small mean squared error, the expectation value of O_i in the output state $\mathcal{E}(\rho)$ for every observable O_i in the family $\{O_i\}$:

$$\mathop{\mathbb{E}}_{\rho \sim \mathcal{D}} |h(\rho, O_i) - \mathrm{tr}(O_i \mathcal{E}(\rho))|^2 \leq \epsilon. \tag{6}$$

Our main result is that this task can be achieved efficiently if O_i is a bounded-degree observable and $\mathcal{D}$ is locally flat. That is, N, the number of times we access $\mathcal{E}$, and the computational complexity of training and prediction scale reasonably with the system size n and the target accuracy ϵ. For example, any generic input product state can be predicted efficiently and accurately. From Eq. (6), it is also easy to see that a small average error can be achieved for any distribution $\mathcal{D}$ that is at most polynomially far away from a locally flat distribution with distance measured by the maximum likelihood ratio.

To prove this result, we observe that the task of learning an unknown quantum process can be reduced to learning unknown states and learning unknown observables. If ρ_ℓ is sampled from distribution $\mathcal{D}$ then, since $\mathcal{E}$ is unknown, $\mathcal{E}(\rho_\ell)$ should be regarded as an unknown quantum state. Suppose that we learn this state; that is, after preparing and measuring $\mathcal{E}(\rho_\ell)$ sufficiently many times we can accurately predict the expectation value $\mathrm{tr}(O_i\mathcal{E}(\rho_\ell))$ for each target observable O_i.

Now note that $\mathrm{tr}(O_i\mathcal{E}(\rho_\ell)) = \mathrm{tr}(\mathcal{E}^\dagger(O_i)\rho_\ell)$, where $\mathcal{E}^\dagger$ is the (Heisenberg-picture) map dual to $\mathcal{E}$. Since $\mathcal{E}^\dagger$ is unknown, $\mathcal{E}^\dagger(O_i)$ should be regarded as an unknown observable. Suppose that we learn this observable; that is, using the dataset $\{\rho_\ell, \mathrm{tr}(\mathcal{E}^\dagger(O_i)\rho_\ell)\}$ as training data, we can predict $\mathrm{tr}(\mathcal{E}^\dagger(O_i)\rho)$ for ρ drawn from $\mathcal{D}$ with a small mean squared error. This achieves the task of learning process $\mathcal{E}$ for state distribution $\mathcal{D}$ and target observable O_i.

Having already shown that arbitrary quantum states can be learned efficiently for the purpose of predicting expectation values of bounded-degree observables and that arbitrary observables can be learned efficiently for any input state distribution that is not superpolynomially far away from a locally flat distribution, we obtain our main result. Since distribution $\mathcal{D}$ is not too far from locally flat, it suffices to learn the low-degree truncated approximation to the unknown operator $\mathcal{E}^\dagger(O_i)$, incurring only a small mean squared error. To predict $\mathrm{tr}(\mathcal{E}^\dagger(O_i)\rho)$, then, it suffices to know only the few-body RDMs of the input state ρ. For any input state ρ, these few-body density matrices can be learned efficiently using classical shadow tomography.

As noted above in the discussion of learning observables, states ρ_ℓ in the training data need not be sampled from $\mathcal{D}$. To learn a low-degree approximation to $\mathcal{E}^\dagger(O_i)$, it suffices to sample from the uniform distribution over product states. Even if we sample only product states during training, we can make accurate predictions for highly entangled input states. We also emphasize again that the unknown process $\mathcal{E}$ is arbitrary. Even if $\mathcal{E}$ has quantum computational complexity exponential in n, we can learn to predict $\mathrm{tr}(O\mathcal{E}(\rho))$ accurately and efficiently for bounded-degree observables O and for any distribution on the input state ρ that is at most polynomially far from some locally flat distribution.

III. ALGORITHM FOR LEARNING AN UNKNOWN QUANTUM PROCESS

Consider an unknown n-qubit quantum process $\mathcal{E}$ (a CPTP map). Suppose that we have obtained a classical dataset by performing N randomized experiments on $\mathcal{E}$. Each experiment prepares a random product state $|\psi^{(\mathrm{in})}\rangle = \bigotimes_{i=1}^n |s_i^{(\mathrm{in})}\rangle$, passes through $\mathcal{E}$, and performs a randomized Pauli measurement [46,47] on the output state. Recall that a randomized Pauli measurement measures each qubit of a state in a random Pauli basis (X, Y, or Z) and produces a measurement outcome of $|\psi^{(\mathrm{out})}\rangle = \bigotimes_{i=1}^n |s_i^{(\mathrm{out})}\rangle$, where $|s_i^{(\mathrm{out})}\rangle \in \mathrm{stab}_1 \triangleq \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\}$. We denote the classical dataset of size N by

$$S_N(\mathcal{E}) \triangleq \left\{ |\psi_\ell^{(\mathrm{in})}\rangle = \bigotimes_{i=1}^n |s_{\ell,i}^{(\mathrm{in})}\rangle \,,\; |\psi_\ell^{(\mathrm{out})}\rangle = \bigotimes_{i=1}^n |s_{\ell,i}^{(\mathrm{out})}\rangle \right\}_{\ell=1}^N, \tag{7}$$

where $|s_{\ell,i}^{(\mathrm{in})}\rangle, |s_{\ell,i}^{(\mathrm{out})}\rangle \in \mathrm{stab}_1$. Each product state is represented classically with $\mathcal{O}(n)$ bits. Hence, the classical dataset $S_N(\mathcal{E})$ is of size $\mathcal{O}(nN)$ bits. The classical dataset can be seen as one way to generalize the notion of classical shadows of quantum states [46] to quantum processes. Our goal is to design an ML algorithm that can learn an approximate model of $\mathcal{E}$ from the classical dataset $S_N(\mathcal{E})$, such that, for a wide range of states ρ and observables O, the ML model can predict a real value $h(\rho, O)$ that is approximately equal to $\mathrm{tr}(O\mathcal{E}(\rho))$.

A. ML algorithm

We are now ready to state the proposed ML algorithm. At a high level, the ML algorithm learns a low-degree approximation to the unknown n-qubit CPTP map $\mathcal{E}$. Despite the simplicity of the ML algorithm, several ideas go into the design of the ML algorithm and the proof of the rigorous performance guarantee. These ideas are presented in Sec. IV below.

Let O be an observable with $\|O\| \le 1$ that is written as a sum of few-body observables, where each qubit is acted on by $\mathcal{O}(1)$ of the few-body observables. We denote the Pauli representation of O as $\sum_{Q\in\{I,X,Y,Z\}^{\otimes n}} a_Q Q$. By the definition of O, there are $\mathcal{O}(n)$ nonzero Pauli coefficients a_Q. We consider a hyperparameter $\tilde\epsilon > 0$; roughly speaking, $\tilde\epsilon$ will scale inverse polynomially in the dataset size N from Eq. (12) below. For every Pauli observable $P \in \{I,X,Y,Z\}^{\otimes n}$ with $|P| \le k = \Theta(\log(1/\epsilon))$, the algorithm computes an empirical estimate for the corresponding Pauli coefficient α_P via

$$\hat{x}_P(O) = \frac{1}{N}\sum_{\ell=1}^N \mathrm{tr}\left(P \bigotimes_{i=1}^n |s_{\ell,i}^{(\mathrm{in})}\rangle\langle s_{\ell,i}^{(\mathrm{in})}|\right) \times \mathrm{tr}\left(O \bigotimes_{i=1}^n (3|s_{\ell,i}^{(\mathrm{out})}\rangle\langle s_{\ell,i}^{(\mathrm{out})}| - I)\right), \tag{8}$$

$$\hat{\alpha}_P(O) = \begin{cases} 3^{|P|}\hat{x}_P(O), & \left(\frac{1}{3}\right)^{|P|} > 2\tilde\epsilon \text{ and } |\hat{x}_P(O)| > 2\cdot 3^{|P|/2}\sqrt{\tilde\epsilon}\sum_{Q:\, a_Q\neq 0}|a_Q|, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The computation of $\hat{x}_P(O)$ and $\hat{\alpha}_P(O)$ can be done classically. The basic idea of $\hat{\alpha}_P(O)$ is to set the coefficient $3^{|P|}\hat{x}_P(O)$ to zero when the influence of Pauli observable P is negligible. Given an n-qubit state ρ, the algorithm outputs

$$h(\rho, O) = \sum_{P:\, |P|\le k} \hat{\alpha}_P(O)\,\mathrm{tr}(P\rho). \tag{10}$$

With a proper implementation, the computational time is $\mathcal{O}(kn^kN)$. Note that, to make predictions, the ML algorithm only needs the k-body reduced density matrices (k-RDMs) of ρ. The k-RDMs of ρ can be efficiently obtained by performing randomized Pauli measurement on ρ and using the classical shadow formalism [46,47]. Except for this step, which may require quantum computation, all other steps of the ML algorithm only require classical computation. Hence, if the k-RDMs of ρ can be computed classically then we have a classical ML algorithm that can predict an arbitrary quantum process $\mathcal{E}$ after learning from data.

B. Rigorous guarantee

To measure the prediction error of the ML model, we consider the average-case prediction performance under an arbitrary n-qubit state distribution $\mathcal{D}$ invariant under single-qubit Clifford gates, which means that the probability distribution $f_{\mathcal{D}}(\rho)$ of sampling a state ρ is equal to $f_{\mathcal{D}}(U\rho U^\dagger)$ of sampling $U\rho U^\dagger$ for any single-qubit Clifford gate U. We call such a distribution locally flat.

Theorem 1 (Learning an unknown quantum process).— Suppose that $\epsilon, \epsilon' = \Theta(1)$ and that there is a training set $S_N(\mathcal{E})$ of size $N = \mathcal{O}(\log n)$ as specified in Eq. (7). With high probability, the ML model can learn a function $h(\rho, O)$ from $S_N(\mathcal{E})$ such that, for any distribution $\mathcal{D}$ over n-qubit states invariant under single-qubit Clifford gates, and for any bounded-degree observable O with $\|O\| \leq 1$,

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2 \leq \epsilon + \max(\|O'\|^2, 1)\epsilon', \tag{11}$$

where O' is the low-degree truncation [of degree $k = \lceil \log_{1.5}(1/\epsilon)\rceil$] of observable O after the Heisenberg evolution under $\mathcal{E}$. The training and prediction time of $h(\rho, O)$ are both polynomial in n. When ϵ is small and $\epsilon' = 0$, the data size N and computational time scale as $2^{\mathcal{O}(\log(1/\epsilon)\log(n))}$.

The detailed theorem statement and the proof of the theorem are given in Appendix E. An interesting aspect of the above theorem is that the states sampled from distribution $\mathcal{D}$ can be highly entangled, even though the training data $S_N(\mathcal{E})$ only contains information about random product states. From the theorem, we can see that if $\|O'\| = \mathcal{O}(1)$ then we only need $\mathcal{O}(\log(n))$ samples to obtain a constant prediction error. Otherwise, $\mathcal{O}(\log(n))$ samples are still enough to guarantee a constant prediction error relative to $\|O'\|^2$. The precise scaling is given as follows. Consider data size

$$N = \log(n) \min\left(2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}},\right.$$
$$\left. 2^{\mathcal{O}[\log(1/\epsilon)\log(n)]}\right). \tag{12}$$

The computational time to learn and predict $h(\rho, O)$ is bounded above by $\mathcal{O}(kn^kN)$ and the prediction error is bounded as

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2 \leq \epsilon + \max(\|O'\|^2, 1)\epsilon'. \tag{13}$$

As we take ϵ' to be zero, we can remove the dependence on the low-degree truncation O'. In this setting, N and computation time both become $2^{\mathcal{O}(\log(1/\epsilon)\log(n))}$, which is polynomial in n if $\epsilon = \Theta(1)$ and is quasipolynomial in n if $\epsilon = 1/\mathrm{poly}(n)$.

For a distribution $\mathcal{D}$ that is not locally flat, we can consider a locally flat distribution $\mathcal{D}^*$ that is closest to $\mathcal{D}$ under the distance defined by the maximum likelihood ratio $\Delta := \sup_\rho [p_{\mathcal{D}}(\rho)/p_{\mathcal{D}^*}(\rho)]$, where $p_{\mathcal{D}}(\rho)$ is the probability density of ρ under $\mathcal{D}$. We can see that the average prediction error under $\mathcal{D}$ satisfies

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2$$
$$\leq \Delta \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^*} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2. \tag{14}$$

Hence, if the distance Δ is at most $\mathrm{poly}(n)$ then the prediction error under $\mathcal{D}$ is small using a quasipolynomial sample complexity and computational time.

IV. PROOF IDEAS

The proof of the rigorous performance guarantee for the proposed ML algorithm consists of five parts. The first two parts presented in Appendices A and B are a detour to establish a few fundamental and useful norm inequalities about Hamiltonians and observables. The latter three parts given in Appendices C, D, and E apply the newly established norm inequalities to three learning tasks. In the following, we present the basic ideas in each part.

A. Improved approximation algorithms for optimizing local Hamiltonians

We begin with a different task, namely, optimizing local Hamiltonians. We are given an n-qubit k-local Hamiltonian

$$H = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|\leq k} \alpha_P P, \tag{15}$$

where $|P|$ is the weight of the Pauli operator P, the number of qubits upon which P acts nontrivially. Our goal is to find a state $|\psi\rangle$ that maximizes or minimizes $\langle\psi| H |\psi\rangle$. This task is related to solving ground states [48,49] when we consider minimizing $\langle\psi| H |\psi\rangle$ and quantum optimization [43,44,50–54] when we consider maximizing $\langle\psi| H |\psi\rangle$.

We give a general randomized approximation algorithm in Appendix A for producing a random product state $|\psi\rangle$ that either approximately minimizes or approximately maximizes a k-local Hamiltonian H with a rigorous upper or lower bound based on the Pauli coefficients α_P of H. The proposed optimization algorithm applies to various classes of Hamiltonians and is inspired by the proofs of Littlewood's 4/3 inequality [55] and the Bohnenblust-Hille inequality [56]. For classes that have been studied previously [41–44], the proposed algorithm yields an improved bound. Our improvement crucially stems from our construction for the random state $|\psi\rangle$. In Refs. [41–43] the authors utilize a random restriction approach, where some random subset of qubits is fixed with some random values and the rest of the qubits are optimized. On the other hand, we utilize a polarization approach, where we replicate each qubit many times, randomly fix all except the last replica, optimize the last replica, and combine using a random-signed averaging. A detailed comparison is given in Appendices A 1 c and A 2.

Two classes of Hamiltonians used in our learning applications are general k-local Hamiltonians and bounded-degree k-local Hamiltonians. A k-local Hamiltonian with degree at most d is a Hermitian operator that can be written as a sum of k-qubit observables, where each qubit is acted on by at most d of the k-qubit observables.

Corollary 1 (Optimizing the general k-local Hamiltonian).—Consider an n-qubit k-local Hamiltonian

$$H = \sum_{P:\,|P|\leq k} \alpha_P P. \tag{16}$$

There is a randomized algorithm that runs in time $\mathcal{O}(n^k)$ and produces either a random maximizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \geq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] + C(k)\left(\sum_{P\neq I} |\alpha_P|^{2k/(k+1)}\right)^{(k+1)/(2k)} \tag{17}$$

or a random minimizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \leq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] - C(k)\left(\sum_{P\neq I} |\alpha_P|^{2k/(k+1)}\right)^{(k+1)/(2k)}, \tag{18}$$

where $C(k) = 1/\exp(\Theta(k \log k))$.

Corollary 2 (Optimizing the bounded-degree k-local Hamiltonian).—Consider an n-qubit k-local Hamiltonian $H = \sum_{P:\,|P|\leq k} \alpha_P P$ with bounded degree d, $|\alpha_P| \leq 1$ for all P, and $k = \mathcal{O}(1)$. There is a randomized algorithm that runs in time $\mathcal{O}(nd)$ and produces either a random maximizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \geq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] + \frac{C}{\sqrt{d}} \sum_{P\neq I} |\alpha_P| \tag{19}$$

or a random minimizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \leq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] - \frac{C}{\sqrt{d}} \sum_{P\neq I} |\alpha_P| \tag{20}$$

for some constant C.

We note that in the above results, we cannot control whether our algorithm outputs an approximate maximizer or minimizer. This caveat stems from the use of polarization, where the random-signed averaging only guarantees improvement in one of the two directions. Modifying our approach to address this issue is an interesting direction for future work.

B. Norm inequalities from approximate optimization algorithms

The bridge that connects the optimization of k-local Hamiltonians and efficient learning of quantum states and processes is a set of norm inequalities. A norm that characterizes the efficiency of learning is the Pauli-p norm, defined as the ℓ_p norm on the Pauli coefficients of a Hamiltonian $H = \sum_P \alpha_P P$,

$$\|H\|_{\text{Pauli},p} \triangleq \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}} |\alpha_P|^p\right)^{1/p}. \tag{21}$$

The rigorous guarantees from the previous section, namely, on finding a state $|\psi\rangle$ whose energy is higher or lower than a Haar-random state by a margin that depends on the Pauli coefficients α_P, give an algorithmic proof that the spectral norm $\|H\|$ and the Pauli coefficients α_P are related. The proof of this relation is given in Appendix B. In particular, for general and bounded-degree k-local Hamiltonians, we can use the rigorous guarantee from the approximation algorithms to obtain the following norm inequalities. Corollary 3 proves the conjecture given in Ref. [45].

Corollary 3 (Norm inequality for the general k-local Hamiltonian).—Given an n-qubit k-local Hamiltonian H, we have

$$\tfrac{1}{3}C(k)\|H\|_{\text{Pauli},2k/(k+1)} \leq \|H\|, \tag{22}$$

where $C(k) = 1/\exp(\Theta(k \log k))$.

Corollary 4 (Norm inequality for the bounded-degree local Hamiltonian).—Given an n-qubit k-local Hamiltonian H with bounded degree d, we have

$$\tfrac{1}{3}C(k,d)\|H\|_{\text{Pauli},1} \leq \|H\|, \tag{23}$$

where $C(k,d) = 1/(\sqrt{d}\exp(\Theta(k \log k)))$.

C. Sample-optimal algorithm for predicting bounded-degree observables

As the first application of the above norm inequalities to learning, we consider the basic problem of predicting many properties of an unknown n-qubit state ρ. Given M observables $O_1, \ldots, O_M$, after performing measurements on multiple copies of ρ, we would like to predict $\text{tr}(O_i\rho)$

to ϵ error for all $i \in \{1, \ldots, M\}$. This is the task known as shadow tomography [46,57,58]. One approach for obtaining practically efficient algorithms for shadow tomography is via the classical shadow formalism [46].

We consider a physically relevant class of observables, where the observable $O_i = \sum_j O_{ij}$ is a sum of few-body observables O_{ij} and each qubit is acted on by $\mathcal{O}(1)$ of the few-body observables. Despite significant recent progress in shadow tomography [8,33,57,59–70], the sample complexity (number of copies of ρ) for predicting this class of observables has not been established. The central challenge is the appearance of the Pauli-1 norm $\|O_i\|_{\text{Pauli},1}$ when characterizing the sample complexity. In particular, one can bound the shadow norm $\|O_i\|_{\text{shadow}}$ [46], which gives an upper bound on the sample complexity in terms of the Pauli-1 norm $\|O_i\|_{\text{Pauli},1}$ up to a constant factor. Using the new norm inequality established in this work, we give a sample-optimal algorithm for predicting bounded-degree observables.

The sample-optimal algorithm is equivalent to performing classical shadow tomography based on randomized Pauli measurements [46,47], and is essentially the ML algorithm given in Sec. III A with a fixed input state. Consider an unknown n-qubit state ρ. After performing N randomized Pauli measurements on N copies of ρ, we have a classical dataset denoted as

$$S_N(\rho) \triangleq \left\{ |\psi_\ell^{(\text{out})}\rangle = \bigotimes_{i=1}^{n} |s_{\ell,i}^{(\text{out})}\rangle \right\}_{\ell=1}^{N}, \tag{24}$$

where $|s_{\ell,i}^{(\text{out})}\rangle \in \text{stab}_1$ is a single-qubit stabilizer state. Given an observable O, the algorithm predicts

$$h(O) = \frac{1}{N}\sum_{\ell=1}^{N} \text{tr}\left(O \bigotimes_{i=1}^{n} (3|s_{\ell,i}^{(\text{out})}\rangle\langle s_{\ell,i}^{(\text{out})}| - I)\right). \tag{25}$$

It is not hard to see that computing $h(O)$ requires only $\mathcal{O}(nN)$ classical computation time. Hence, as we show later that $N = \mathcal{O}(\log(n)/\epsilon^2)$, the learning algorithm is very efficient. Using the norm inequality for bounded-degree local Hamiltonian $\|H\|_{\text{Pauli},1} \leq C\|H\|$ for a constant C in Corollary 4, and the classical shadow formalism [46,47], we obtain the following performance guarantee.

Theorem 2 (Sample complexity upper bound).—Consider an unknown n-qubit state ρ and any n-qubit observable $O_1, \ldots, O_M$ with $\|O_i\| \leq B_\infty$. Suppose that each observable O_i is a sum of few-body observables, where each qubit is acted on by $\mathcal{O}(1)$ of the few-body observables. Using a classical dataset $S_N(\rho)$ of size

$$N = \mathcal{O}\left(\frac{\log(\min(M,n))B_\infty^2}{\epsilon^2}\right), \tag{26}$$

we have $|h(O_i) - \text{tr}(O_i\rho)| \leq \epsilon$ for all $i \in \{1, \ldots, M\}$ with high probability. The constant factor in the $\mathcal{O}(\cdot)$ notation above scales polynomially in the degree and exponentially in the locality of the observables.

The following theorem shows that the above algorithm achieves the optimal sample complexity of any algorithm that can perform collective measurement on many copies of ρ.

Theorem 3 (Sample complexity lower bound).—Consider the following task. There is an unknown n-qubit state ρ, and we are given M observables $O_1, \ldots, O_M$ with $\max_i\|O_i\| \leq B_\infty$. Each observable O_i is a sum of few-body observables, where every qubit is acted on by $\mathcal{O}(1)$ of the few-body observables. We would like to estimate $\text{tr}(O_i\rho)$ to ϵ error for all $i \in [M]$ with high probability by performing arbitrary collective measurements on N copies of ρ. The number of copies N must be at least

$$N = \Omega\left(\frac{\log(\min(M,n))B_\infty^2}{\epsilon^2}\right) \tag{27}$$

for any algorithm to succeed in this task.

The detailed proofs of the sample complexities stated in the above theorems are given in Appendix C.

D. Efficient algorithms for learning an unknown observable from log(n) samples

As a second learning application of the norm inequalities, we consider the task of learning an unknown n-qubit observable $O^{(\text{unk})} = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \alpha_P P$. We can think of this unknown observable as $\mathcal{E}^\dagger(O)$, i.e., the observable O after Heisenberg evolution under the unknown process $\mathcal{E}$. Suppose that we are given a training dataset of $\{\rho_\ell, \text{tr}(O^{(\text{unk})}\rho_\ell)\}_{\ell=1}^{N}$, where ρ_ℓ is sampled from an arbitrary distribution $\mathcal{D}$ over n-qubit states that is invariant under single-qubit Clifford gates. Given an integer $k > 0$, we define the weight-k truncation of $O^{(\text{unk})}$ to be the Hermitian operator

$$O^{(\text{unk},k)} \triangleq \sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|\leq k} \alpha_P P, \tag{28}$$

where $|P|$ is the number of qubits upon which P acts nontrivially. For a small k, we can think of $O^{(\text{unk},k)}$ as a low-weight approximation of the unknown observable $O^{(\text{unk})}$. By definition, $O^{(\text{unk},k)}$ is a k-local Hamiltonian; hence, the norm inequality in Corollary 3 shows that

$$\begin{aligned}\frac{1}{3}C(k)\|O^{(\text{unk},k)}\|_{\text{Pauli},2k/(k+1)} \\ = \frac{1}{3}C(k)\left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:|P|\leq k} |\alpha_P|^r\right)^{1/r} \leq \|O^{(\text{unk},k)}\|,\end{aligned} \tag{29}$$

where $r = 2k/(k+1) \in [1, 2)$. An ℓ_r-norm bound ($r < 2$) on the Pauli coefficients implies that we can remove most of the small Pauli coefficients without incurring too much change under the ℓ_2 norm. As an example, consider an M-dimensional vector x with $\|x\|_r \leq 1$. Given $\widetilde{\epsilon} > 0$, let $\widetilde{x}$ be the M-dimensional vector with $\widetilde{x}_i = x_i$ if $|x_i| > \widetilde{\epsilon}$ and $\widetilde{x}_i = 0$ if $|x_i| \leq \widetilde{\epsilon}$. We have

$$\begin{aligned}\|x-\widetilde{x}\|_2^2 &= \sum_{i:\,|x_i|\leq\widetilde{\epsilon}} |x_i|^2 \leq \widetilde{\epsilon}^{2-r} \sum_{i:\,|x_i|\leq\widetilde{\epsilon}} |x_i|^r \\ &\leq \widetilde{\epsilon}^{2-r} \sum_i |x_i|^r \leq \widetilde{\epsilon}^{2-r}. \end{aligned} \tag{30}$$

In Appendix D 1, we show that the average error (both the mean squared error and the mean absolute error) is characterized by the ℓ_2 norm. Hence, Eq. (29) implies that we can set most of the Pauli coefficients in $O^{(\mathrm{unk},k)}$ to zero without incurring too much error on average.

Using the above reasoning, learning the low-weight truncation $O^{(\mathrm{unk},k)}$ amounts to learning the large Pauli coefficients of $O^{(\mathrm{unk},k)}$ and setting all small Pauli coefficients to zero. This ensures that the learning can be done very efficiently. This approach is presented in Appendix D 2 with the main result stated in Lemma 18. It is inspired by the learning algorithm of Ref [71] that achieves a logarithmic sample complexity for learning classical low-degree functions.

The last step in the proof is to argue that the low-weight truncation $O^{(\mathrm{unk},k)}$ is a good surrogate for the unknown observable $O^{(\mathrm{unk})}$ when the goal is to predict $\mathrm{tr}(O^{(\mathrm{unk})}\rho)$. The key insight here is that, for distributions $\mathcal{D}$ that are invariant under single-Clifford gates, the contribution of any Pauli term P in $O^{(\mathrm{unk})}$ to $\mathbb{E}_{\rho\sim\mathcal{D}}[\mathrm{tr}(O^{(\mathrm{unk})}\rho)^2]$ is *exponentially decaying* in weight $|P|$. This allows us to prove that $\mathbb{E}_{\rho\sim\mathcal{D}}[\mathrm{tr}((O^{(\mathrm{unk})} - O^{(\mathrm{unk},k)})\rho)^2]$ is small.

Putting these ingredients together, we arrive at the following theorem. As stated in the theorem, the learning algorithm is computationally efficient.

Theorem 4 (Learning an unknown observable).—Suppose that $\epsilon, \epsilon', \delta > 0$. Let $k = \lceil \log_{1.5}(1/\epsilon) \rceil$ and $r = 2k/(k+1) \in [1,2)$. From training data $\{\rho_\ell, \mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell)\}_{\ell=1}^N$ of size

$$\begin{aligned} N = \log(n/\delta) \min \Big(& 2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}, \\ & 2^{\mathcal{O}[\log(1/\epsilon)\log(n)]} \Big), \end{aligned} \tag{31}$$

where ρ_ℓ is sampled from $\mathcal{D}$, we can learn a function $h(\rho)$ such that

$$\begin{aligned} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 &\leq (\epsilon+\epsilon')\|O^{(\mathrm{unk})}\|^2 \\ &\quad + \epsilon' \|O^{(\mathrm{unk},k)}\|^r \|O^{(\mathrm{unk})}\|^{2-r} \end{aligned} \tag{32}$$

with probability at least $1-\delta$. The training and prediction times of $h(\rho)$ are $\mathcal{O}(Nn^k)$.

The factor of $\|O^{(\mathrm{unk})}\|^2$ in the prediction error is the natural scale of the squared error. From the theorem, we can see that we only need $\mathcal{O}(\log(n))$ samples to obtain a constant prediction error relative to $\|O^{(\mathrm{unk})}\|^2 + \|O^{(\mathrm{unk},k)}\|^r\|O^{(\mathrm{unk})}\|^{2-r}$. The proof of the theorem and the detailed description of the ML algorithm are given in Appendix D.

E. Learning an unknown quantum process

The ML algorithm for learning an unknown n-qubit quantum process $\mathcal{E}$ is essentially the combination of the two learning applications described above with a few modifications. At a high level, we consider the following. There is an n-qubit state ρ sampled from an unknown distribution $\mathcal{D}$, as well as an observable O that can be written as a sum of few-body observables, where each qubit is acted on by a constant number of the few-body observables. In the first stage, we use the sample-optimal algorithm for predicting the bounded-degree observable O, where $\mathcal{E}(\rho_\ell)$ is an unknown quantum state, thus transforming the classical dataset $S_N(\mathcal{E})$ in Eq. (7) into a dataset,

$$\{\rho_\ell \triangleq |\psi_\ell^{(\mathrm{in})}\rangle\langle\psi_\ell^{(\mathrm{in})}|,\ \mathrm{tr}(O\mathcal{E}(\rho_\ell))\}_{\ell=1}^N, \tag{33}$$

that maps quantum states to real numbers. In the second stage, we apply the efficient algorithm for learning an unknown observable $O^{(\mathrm{unk})} = \mathcal{E}^\dagger(O)$, regarding Eq. (33) as the training data for this task, thus predicting $\mathrm{tr}(\mathcal{E}^\dagger(O)\rho) = \mathrm{tr}(O\mathcal{E}(\rho))$ for state ρ drawn from distribution $\mathcal{D}$. Because both stages of the algorithm run in time polynomial in n, the overall runtime for this procedure is polynomial in n.

In our actual proofs, there are a few deviations from the above high-level design, stemming from the fact that the input states ρ_ℓ are tensor products of random single-qubit stabilizer states. This specific setting allows a few simplifications to be made. With the simplifications, we can remove an additive factor of ϵ' in the prediction error. Furthermore, a surprising fact is that learning from random product states is sufficient to predict highly entangled states sampled from any distribution $\mathcal{D}$ invariant under single-qubit Clifford unitaries. This surprising fact is a result of the characterization of the prediction error given in Lemma 14 based on a modified purity on subsystems of an input quantum state $\rho \sim \mathcal{D}$.

By combining the five parts, we can establish Theorem 1, the precise sample complexity scaling in Eq. (12), and the prediction error bound in Eq. (13). The full proof is given in Appendix E.

V. NUMERICAL EXPERIMENTS

We have conducted numerical experiments to assess the performance of ML models in learning the dynamics of several physical systems. The results corroborate our theoretical claims that long-time evolution over a many-body system can be learned efficiently. While our theorem only guarantees good performance for randomly sampled input states, we also find that the ML models work very well for structured input states that could be of practical interest. The source code is available from a public GitHub repository [72]. We note that all prior tomographic protocols that can learn an arbitrary quantum process require a sample complexity that scales exponentially in n. The strong numerical performance demonstrated here raises the hope that the synthesis of existing tomographic techniques and the low-weight truncation proposed in this work will enable a more powerful ML method for predicting arbitrary quantum processes.

We focus on training ML models to predict output state properties after the time dynamics of one-dimensional (1D) n-spin XY and Ising chains with homogeneous or disordered Z fields. Let H be the many-body Hamiltonian. The quantum process $\mathcal{E}$ is given by $\mathcal{E}(\rho) = e^{-itH}\rho e^{itH}$ for a significantly long evolution time $t = 10^6$. We consider the ML models described by Eq. (10). While we utilize the very simple sparsity-enforcing strategy of setting small values to zero to prove Theorem 1, the standard sparsity-enforcing approach is through ℓ_1 regularization [73]. A detailed description of applying ℓ_1 regularization to enforce sparsity in $\alpha_P(O)$ is given in Appendix F. We find the best hyperparameters using fourfold cross-validation to minimize the root-mean-square error (RMSE) and report the predictions on a test set.

Figure 2 considers the performance for predicting the expectation of the Pauli-Z operator Z_i on the output state for randomly sampled product input states not in the training data. Figure 2(a) illustrates the many-body Hamiltonian H. Figure 2(b) shows the dependence of the error on the training set size N. We can clearly see that, as the training set size N increases, the prediction error notably decreases. This observation confirms our theoretical claim that long-time quantum dynamics could be efficiently learned. In Fig. 2(c), we consider how the evolution time t affects prediction performance. From the figure, we can see that, even when we exponentially increase t, the prediction performance remains similar. This matches with our theorem stating that no matter what the quantum process $\mathcal{E}$ is, even if $\mathcal{E}$ is an exponentially long-time dynamics, the ML model can still predict accurately and efficiently. In Fig. 2(d), we consider the dependence on the system size n. As n increases linearly, the Hilbert space dimension 2^n grows exponentially. Despite the exponential growth, even for 50-spin systems, the ML model still predicts well. This matches with the logarithmic scaling on n given in Theorem 1.

In Fig. 3, we consider predicting properties of the final state after long-time dynamics for a highly structured input

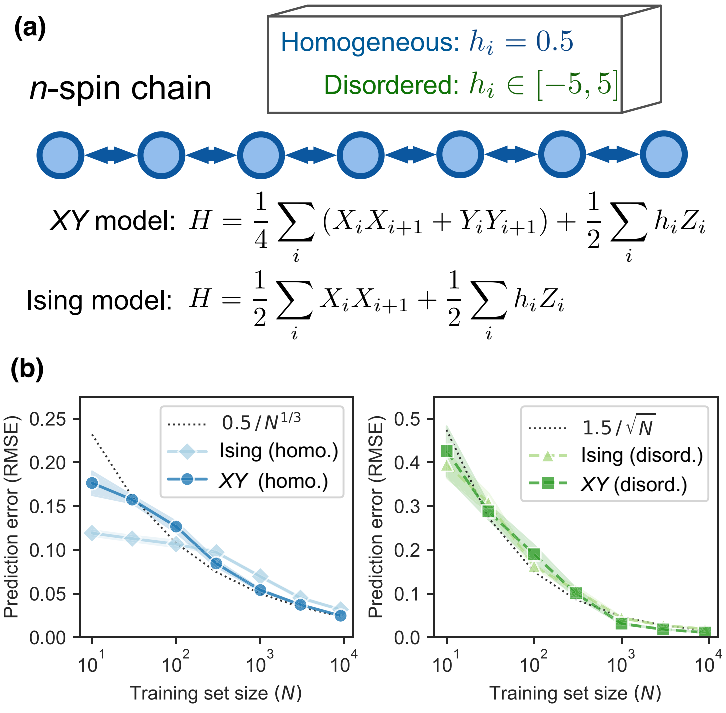

FIG. 2. Prediction performance of ML models for learning $\mathcal{E}(\rho) = e^{-itH}\rho e^{itH}$ for a large time t. (a) Hamiltonians. We consider an XY or Ising model with a homogeneous or disordered Z field on an n-spin open chain. (b) Error scaling with training set size (N). We show the root-mean-square error (RMSE) for predicting the Pauli-Z operator Z_i on the output state $\mathcal{E}(\rho)$ for random product states ρ. (c),(d) Error scaling with evolution time (t) and system size (n). Panel (d) shows the RMSE for the XY model with a homogeneous Z field. The prediction error remains similar as we exponentially increase t and the Hilbert space dimension 2^n.

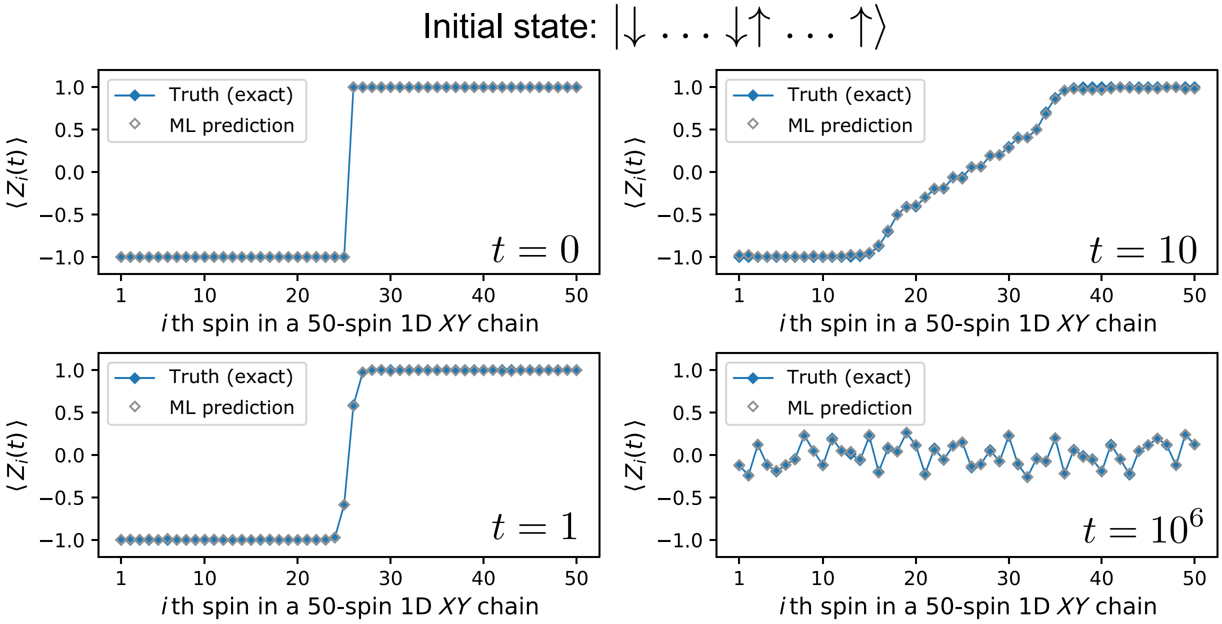

FIG. 3. Visualization of the ML model's prediction for an initial state $\rho = |\psi\rangle\langle\psi|$ with a domain wall. We consider the 1D 50-spin XY chain with a homogeneous Z field. We show the expectation value of $Z_i(t) = e^{itH} Z_i e^{-itH}$ for all the 50 spins on the initial state $|\psi\rangle = |\downarrow \cdots \downarrow\uparrow \cdots \uparrow\rangle$. The ML model is trained on 10 000 random product states. We see that the ML model performs accurately for a significantly large range of time t.

product state

$$|\psi\rangle = |\downarrow \cdots \downarrow\uparrow \cdots \uparrow\rangle, \tag{34}$$

which has a single domain wall in the middle. We focus on predicting the expected value for $Z_i(t) = e^{itH} Z_i e^{-itH}$ on every spin in the 1D 50-spin XY chain with a homogeneous Z field $h_i = 0.5$ and consider evolution time t from 0 to 10^6. We train the ML model using $N = 10\,000$ random input product states. We can see that the ML model predicts very well for this highly structured product state. The collapse of the domain wall is accurately predicted by the ML model despite only seeing outcomes from random unstructured product states. This numerical experiment suggests that the performance of the ML model goes beyond Theorem 1, which only guarantees accurate prediction on average.

Theorem 1 states that the ML model can predict well on highly entangled input states after learning only from random product state inputs. We test this claim in Fig. 4 by considering an entangled input state

$$|\psi_e\rangle = \sum_{\substack{s\in\{\leftarrow,\rightarrow\}^{n/2} \\ \text{with an even \# of} \rightarrow}} \frac{1}{\sqrt{2^{(n/2)-1}}} |s\rangle$$
$$\times \otimes |\rightarrow\downarrow\leftarrow\uparrow\rightarrow\downarrow\leftarrow\uparrow \cdots\rangle. \tag{35}$$

The left $n/2$ spins of state $|\psi_e\rangle$ exhibit Greenberger-Horne-Zeilinger (GHZ)-like entanglement, which requires a linear-depth 1D quantum circuit to prepare. The right $n/2$ spins of $|\psi_e\rangle$ form a product state with spins rotating clockwise from left to right. Combining the left and right spins, state $|\psi_e\rangle$ cannot be generated by a short-depth 1D quantum circuit. We can see that, for this entangled input state, the ML model trained on random product states still predicts very well across a broad range of the evolution time t.

VI. OUTLOOK

The theorem established in this work shows that learning to predict a complex quantum process can be achieved with computationally efficient ML algorithms. Once we have obtained training data by accessing the unknown process $\mathcal{E}$ sufficiently many times, the proposed ML algorithm is entirely classical except for the step of obtaining the RDM of the input state ρ, which may require quantum computation. This algorithm is reminiscent of recent proposals for quantum ML based on kernel methods [2,3,29], in particular the projected quantum kernel [29]. This result highlights the potential for using hybrid quantum-classical ML algorithms to learn to model exotic quantum dynamics occurring in nature.

The results presented in this work also have implications for several previously studied problems. Prior works [7,11, 12] have proposed to train quantum ML models on a given quantum process with the hope that the learned model can be faster than the process itself. Our proof that one can always train an ML model that runs in quasipolynomial

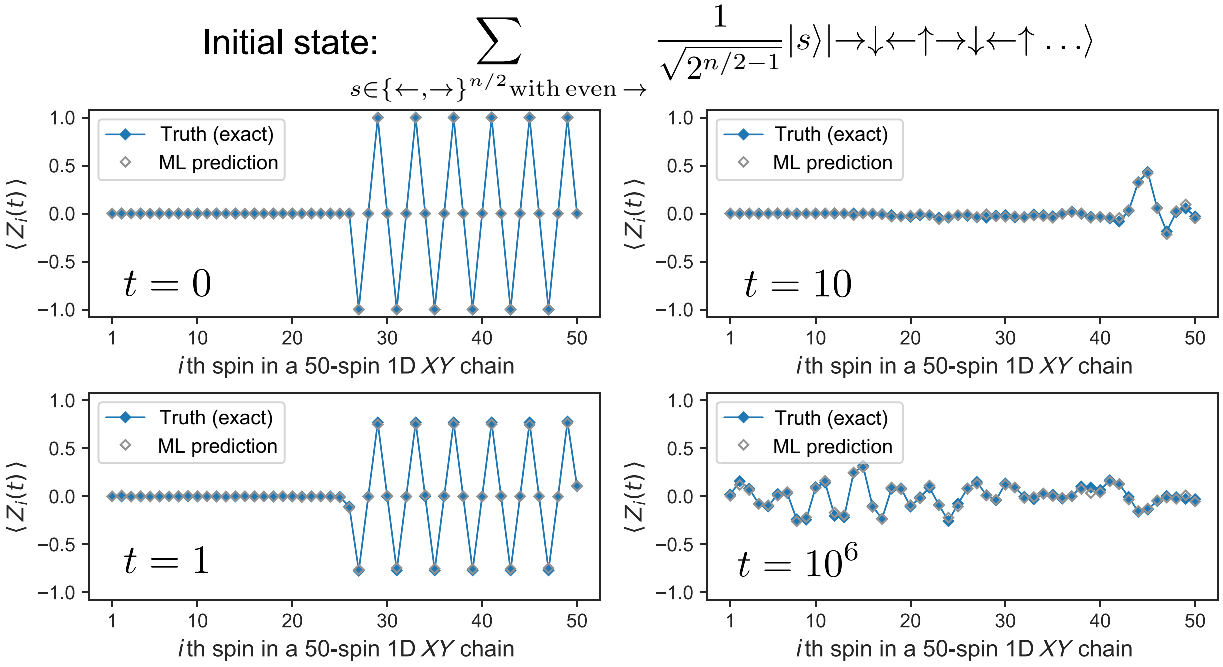

FIG. 4. Visualization of the ML model's prediction for a highly entangled initial state $\rho = |\psi\rangle\langle\psi|$. We consider the expected value of $Z_i(t) = e^{itH} Z_i e^{-itH}$, where H corresponds to the 1D 50-spin XY chain with a homogeneous Z field. The initial state $|\psi\rangle$ has a GHZ-like entanglement over the left-half chain and is a product state with spins rotating clockwise over the right-half chain. To prepare $|\psi\rangle$ with 1D circuits, a depth of at least $\Omega(n)$ is required. Even though the ML model is trained only on random product states (a total of $N = 10\,000$), it still performs accurately in predicting the highly entangled state over a wide range of evolution time t.

time, even for exponential-time quantum dynamics, provides rigorous support for such a hope. When the few-body RDMs of the input state ρ are hard to compute classically, the proposed ML algorithm can be seen as a variant of the projected quantum kernel method [29]. When the few-body RDMs of the input state ρ are easy to compute classically, the proposed ML model can efficiently run on a classical computer. Hence, this result provides a rigorous foundation for empirical works using classical ML to learn and simulate quantum dynamics [27,74–76]. When $\mathcal{E}$ is a parameterized quantum circuit U_θ, such as a quantum neural network [3,4,6,9,29,37], the existence of a classical ML model that can efficiently predict the output of U_θ implies that the function $\mathrm{tr}(OU_\theta \rho U_\theta^\dagger)$ is easy to represent and learn on a classical computer. This finding shows that quantum circuits do not have strong representational power for various distributions over quantum state input ρ with easy-to-compute RDMs.

Several open problems remain to be answered. While we focus only on locally flat distributions $\mathcal{D}$, we believe that efficient ML algorithms also exist for other general classes of distributions. An important open problem is hence the following: can we obtain computationally efficient learning algorithms for any "smooth" distribution over quantum state space? If not, how general can the class of distributions be? Similar questions can be asked about the class of observables that we predict. For what general classes of observables O can one predict efficiently, in terms of both sample size and computation time? This problem is closely related to the problem of when shadow tomography [57,58,77] can be made computationally efficient. Other important questions include the following. If we restrict the quantum process $\mathcal{E}$ to be generated in polynomial time, can we obtain improved efficiency? What efficiency guarantees apply to fermionic or bosonic systems? A better understanding of these problems would illuminate the ultimate power of classical and quantum ML algorithms for learning about physical dynamics.

ACKNOWLEDGMENTS

The authors thank Victor V. Albert, Chi-Fang (Anthony) Chen, Bryan Clark, Richard Kueng, and Spiros Michalakis for valuable input and inspiring discussions. After learning about our proof of the quantum Bohnenblust-Hille inequality, Alexander Volberg and Haonan Zhang found a very different proof with a better Bohnenblust-Hille constant. We thank them for sharing their results with us. H.H. is supported by a Google PhD fellowship. S.C. is supported by NSF Grant No. 2103300. J.P. acknowledges support from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (DE-NA0003525, DE-SC0020290), the U.S. Department of Energy, Office of Science, National Quantum Information

Science Research Centers, Quantum Systems Accelerator, and the National Science Foundation (PHY-1733907). The Institute for Quantum Information and Matter is an NSF Physics Frontiers Center.

APPENDIX A: OPTIMIZING A k-LOCAL HAMILTONIAN WITH RANDOM PRODUCT STATES

While our goal is to design a good ML algorithm with low sample complexity, this appendix is a detour to a different task on the optimization of a k-local Hamiltonian. We present an improved approximation algorithm for optimizing any k-local Hamiltonian. The central result in this detour will become useful for showing the low sample complexity of several ML algorithms.

1. Task description and main theorem

Task 1 (Optimizing a quantum Hamiltonian).—Given $n, k \geq 1$ and an n-qubit k-local Hamiltonian

$$H = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}:\, |P| \leq k} \alpha_P P, \tag{A1}$$

where $|P|$ is the number of nonidentity components in P, find a state $|\psi\rangle$ that maximizes or minimizes $\langle\psi| H |\psi\rangle$.

The task given above is related to solving ground states [48,49] when we consider minimizing $\langle\psi| H |\psi\rangle$ and quantum optimization [43,44,50–54] when we consider maximizing $\langle\psi| H |\psi\rangle$. The maximization and minimization are often the same problem since maximizing $\langle\psi| H |\psi\rangle$ is the same as minimizing $\langle\psi| (-H) |\psi\rangle$. Without further constraints, even for $k = 2$, finding the optimal state $|\psi^*\rangle$ maximizing $\langle\psi| H |\psi\rangle$ is known to be QMA hard [78]; hence, it is expected to have no polynomial-time algorithm even on a quantum computer. Most existing works consider deterministic or randomized constructions of $|\psi\rangle$ with rigorous upper and lower bound guarantees on $\langle\psi| H |\psi\rangle$ for minimization and maximization. Some of these lower bounds [52–54] are based on the optimal value $\mathrm{OPT} = \sup_{|\psi\rangle} \langle\psi| H |\psi\rangle$, while some [43,44,51] are based on the Pauli coefficients α_P.

a. Definition of expansion

In this section, we present a random product state construction for the optimization problem, where the rigorous upper or lower bound is based on the Pauli coefficients α_P and the expansion property defined below. The expansion property is defined for any Hamiltonian H.

Definition 1 (Expansion property).—Given an n-qubit Hamiltonian $H = \sum_P \alpha_P P$, we say that H has an expansion coefficient c_e and expansion dimension d_e if, for any $\Upsilon \subseteq \{1, \ldots, n\}$ with $|\Upsilon| = d_e$,

$$\sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \mathbb{1}[\alpha_P \neq 0 \text{ and } (\Upsilon \subseteq \mathsf{dom}(P) \text{ or } \mathsf{dom}(P) \subseteq \Upsilon)] \leq c_e, \tag{A2}$$

where $\mathsf{dom}(P)$ is the set of qubits that P acts nontrivially on.

The expansion property captures the connectivity of the Hamiltonian. We give two examples, the general k-local Hamiltonian and the geometrically local Hamiltonian, to provide more intuition on the expansion property.

Fact 1 (Expansion property for a general k-local Hamiltonian).—Any Hamiltonian given by a sum of k-qubit observables has expansion coefficient 4^k and expansion dimension k.

Proof.—Let $H = \sum_P \alpha_P P$. All the Pauli observables P with nonzero α_P act on at most k qubits. For any Υ with $|\Upsilon| = k$, all the Pauli observables with nonzero α_P must have a domain contained in Υ. There are at most 4^k such Pauli observables. Hence, the claim follows. ■

Fact 2 (Expansion property for a bounded-degree k-local Hamiltonian).—Any Hamiltonian given by a sum of k-qubit observables $H = \sum_j h_j$, where each qubit is acted on by at most d of the k-qubit observables h_j, has expansion coefficient $c_e = 4^k d$ and expansion dimension $d_e = 1$.

Proof.—For every Υ with $|\Upsilon|$, $\Upsilon = \{i\}$ for some qubit i. For each qubit i (corresponding to $\Upsilon = \{i\}$), we have at most d k-qubit observables acting on i. Each of the k-qubit observables can be expanded into at most 4^k Pauli terms. Hence we can set $c_e = 4^k d$ and $d_e = 1$. ■

Fact 3 (Expansion property for a geometrically local Hamiltonian).—Any Hamiltonian given by a sum of geometrically local observables has expansion coefficient $c_e = \mathcal{O}(1)$ and expansion dimension 1.

Proof.—For a geometrically local Hamiltonian $H = \sum_P \alpha_P P$, each qubit i is acted on by at most a constant number $c_i = \mathcal{O}(1)$ of P with nonzero α_P. Hence, for any qubit i, $\sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \mathbb{1}[\alpha_P \neq 0 \text{ and } (\Upsilon \subseteq \mathsf{dom}(P) \text{ or } \mathsf{dom}(P) \subseteq \Upsilon)] = c_i$. Thus, we can set $d_e = 1$ and $c_e = \max_i c_i = \mathcal{O}(1)$. ■

b. Main theorem

With the expansion property defined, we can state the rigorous guarantee on the performance of the proposed randomized approximation algorithm on optimizing an n-qubit k-local Hamiltonian H. We compare with the average energy $\mathbb{E}_{|\phi\rangle:\,\mathrm{Haar}}[\langle\phi| H |\phi\rangle] = \alpha_I$ over the Haar random state. The randomized approximation algorithm uses an optimization over a single-variable polynomial that guarantees improvement in at least one direction (minimization or maximization).

Theorem 5 (Random product states for optimizing a k-local Hamiltonian).—Consider an n-qubit k-local Hamiltonian $H=\sum_{P:\,|P|\leq k}\alpha_P P$ with expansion coefficient c_e and dimension d_e. Let $r=2d_e/(d_e+1)\in[1,2)$ and $\mathsf{nnz}(H)\triangleq|\{P\colon\alpha_P\neq 0\}|$. There is a randomized algorithm that runs in time $\mathcal{O}(nk+\mathsf{nnz}(H)2^k)$ and produces either a random maximizing state $|\psi\rangle=|\psi_1\rangle\otimes\cdots\otimes|\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi|\,H\,|\psi\rangle]\geq\mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]+C(c_e,d_e,k)\left(\sum_{P\neq I}|\alpha_P|^r\right)^{1/r}\tag{A3}$$

or a random minimizing state $|\psi\rangle=|\psi_1\rangle\otimes\cdots\otimes|\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi|\,H\,|\psi\rangle]\leq\mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]-C(c_e,d_e,k)\left(\sum_{P\neq I}|\alpha_P|^r\right)^{1/r}.\tag{A4}$$

The constant $C(c_e,d_e,k)$ is given by

$$C(c_e,d_e,k)=\frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)}k^{k+1.5+1/r}(\sqrt{6}+2\sqrt{3})^k}=\Theta_k\left(\frac{1}{c_e^{1/(2d_e)}}\right),\tag{A5}$$

where Θ_k considers the asymptotic scaling when k is a constant.

Some observations can be made. First, the improvement over Haar random states in Theorem 5 becomes larger when the expansion coefficient c_e is smaller. Second, $(\sum_{P\neq I}|\alpha_P|^r)^{1/r}$ is the ℓ_r norm on the nonidentity Pauli coefficients, so by monotonicity of ℓ_r norms, $(\sum_{P\neq I}|\alpha_P|^r)^{1/r}$ becomes smaller as r becomes larger (corresponding to larger d_e). Hence, the improvement is greater for smaller expansion dimension d_e. In particular, it is helpful to contrast Eqs. (A3) and (A4) with the following basic estimate corresponding to $r=2$ that holds regardless of c_e,d_e,k:

$$\sup_{|\psi\rangle}\left|\langle\psi|\,H\,|\psi\rangle-\mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]\right|\geq\left(\sum_{P\neq I}|\alpha_P|^2\right)^{1/2}.\tag{A6}$$

This holds for any Hamiltonian $H=\sum_P\alpha_P P$ because $(\sum_{P\neq I}|\alpha_P|^2)^{1/2}=1/2^{n/2}\|H-\alpha_I I\|_F\leq\|H-\alpha_I I\|_\infty$, where $\|\cdot\|$ denotes the spectral norm and $\alpha_I=\mathbb{E}_{|\psi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]$. This basic estimate shows that we can always find a state that improves by at least the ℓ_2 norm of α_P, although the optimization process can be computationally hard.

c. An alternative version of the main theorem

By following the proof of Theorem 5 and replacing the use of Corollary 9 by Lemma 5, we can establish the following alternative theorem statement that does not utilize the expansion property.

Theorem 6 (Random product states for optimizing a k-local Hamiltonian; alternative).—Consider an n-qubit k-local Hamiltonian $H=\sum_{P:\,|P|\leq k}\alpha_P P$ with $k=\mathcal{O}(1)$. Let $\mathsf{nnz}(H)\triangleq|\{P\colon\alpha_P\neq 0\}|$. There is a randomized algorithm that runs in time $\mathcal{O}(nk+\mathsf{nnz}(H)2^k)$ and produces a random state $|\psi\rangle=|\psi_1\rangle\otimes\cdots\otimes|\psi_n\rangle$ satisfying

$$\left|\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi|\,H\,|\psi\rangle]-\mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]\right|\geq D\sum_{i\in[n],p\in\{X,Y,Z\}}\sqrt{\sum_{P:\,P_i=p}\alpha_P^2}\tag{A7}$$

for some constant D.

We can compare the above theorem with a closely related result in Ref. [43]. The following is a restatement of the approximation guarantee from Theorem 2 and Lemma 3 of Ref. [43], which is a corollary of a powerful result in Boolean function analysis [41,42] relating the maximum influence and the ability to sample a bitstring from the Boolean hypercube with a large magnitude in the function value. We can define the influence of qubit i under Pauli matrix $p\in\{X,Y,Z\}$ as $I(i,p)=\sum_{P:\,P_i=p}\alpha_P^2$.

Theorem 7 (Approximation guarantee from Ref. [43] for optimizing a k-local Hamiltonian).—Given an n-qubit k-local Hamiltonian $H=\sum_{P:\,|P|\leq k}\alpha_P P$ with $k=\mathcal{O}(1)$, there is a polynomial-time randomized algorithm that produces a random state $|\psi\rangle=|\psi_1\rangle\otimes\cdots\otimes|\psi_n\rangle$ satisfying

$$\left|\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi|\,H\,|\psi\rangle]-\mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}}[\langle\phi|\,H\,|\phi\rangle]\right|\geq D\sum_{i\in[n],p\in\{X,Y,Z\}}\frac{\sum_{P:\,P_i=p}\alpha_P^2}{\max_{j,q}\sqrt{\sum_{P:\,P_j=q}\alpha_P^2}}\tag{A8}$$

for some constant D.

The guarantee from Ref. [43] is asymptotically optimal when the influence $I(i,p)$ is of a similar magnitude for different qubits i and Pauli matrices p. However, the approximation guarantee can be far from optimal when there is a large variation in the influence $I(i,p)$ over different qubits i,p. As an example, consider a 1D n-qubit nearest-neighbor chain, where $|\alpha_P|=1$ for only a constant number of Pauli observables P and $|\alpha_P|=1/\sqrt{n}$ for the rest of the Pauli observables. The improvements over the Haar random state by our algorithm and the algorithm in

Ref. [43] are respectively given by

$$\Theta\left(\sum_{i\in[n],p\in\{X,Y,Z\}} \sqrt{\sum_{P:\,P_i=p} \alpha_P^2} \right) = \Theta(\sqrt{n}), \tag{A9}$$

$$\Theta\left(\sum_{i\in[n],p\in\{X,Y,Z\}} \frac{\sum_{P:\,P_i=p} \alpha_P^2}{\max_{j,q} \sqrt{\sum_{P:\,P_j=q} \alpha_P^2}} \right) = \Theta(1). \tag{A10}$$

Hence, when there is large variation in the influence, our guarantee improves over that of Ref. [43]. For our machine-learning applications, the removal of the dependence on the maximum influence is central. By removing the ratio $\sqrt{I(i,p)}/\max_{j,q}\sqrt{I(j,q)}$, we can obtain the ℓ_r norm dependence for an $r < 2$, as given in Theorem 5. We will later see that having the ℓ_r-norm bound (for $r < 2$) allows a substantial reduction in the sample complexity in training machine-learning models for predicting properties.

We do want to mention that the improvement comes at a cost of a slightly worse dependence on $k = \mathcal{O}(1)$. In Theorem 7 from Ref. [43] based on Boolean function analysis [41,42], the dependence on D is $1/2^{\Theta(k)}$. However, our result in Theorem 6 is $D = 1/2^{\Theta(k \log k)}$. This difference stems from the construction for the random state $|\psi\rangle$. In Refs. [41–43] the authors utilize a random restriction approach, where some random subset of variables is fixed with some random values and the rest of the variables are optimized. On the other hand, we utilize a polarization approach, where we replicate each variable many times, randomly fix all except the last replica, optimize the last replica, and combine using a random-signed averaging.

2. Corollaries of the main theorem

Here, we consider how the main theorem applies to certain classes of k-local Hamiltonians and discuss the relations of the corollaries to related works.

a. Optimizing arbitrary *k*-local Hamiltonians

The first corollary considers a general k-local Hamiltonian $H = \sum_{P:\,|P|\leq k} \alpha_P P$. We can combine Fact 1 and the main theorem to obtain the following corollary.

Corollary 5 (Optimizing an arbitrary k-local Hamiltonian).—Given an n-qubit k-local Hamiltonian $H = \sum_{P:\,|P|\leq k} \alpha_P P$, there is a randomized algorithm that runs in time $\mathcal{O}(n^k)$ and produces a random product state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ with

$$\left| \mathop{\mathbb{E}}_{|\psi\rangle} [\langle\psi| H |\psi\rangle] - \mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}} [\langle\phi| H |\phi\rangle] \right| \geq C(k) \left(\sum_{P\neq I} |\alpha_P|^{2k/(k+1)} \right)^{(k+1)/(2k)}, \tag{A11}$$

where $C(k) = \sqrt{2(k!)}/[2k^{k+1.5+(k+1)/(2k)}(\sqrt{6}+2\sqrt{3})^k]$.

For $k = 2$, we have $2k/(k+1) = 4/3$ and the above result resembles Littlewood's $4/3$ inequality. Recall that Littlewood's $4/3$ inequality states that, given $\{\beta_{i,j} \in \mathbb{C}\}_{i,j}$,

$$\sup \left\{ \left| \sum_{i,j} \beta_{i,j} x_i^{(1)} x_j^{(2)} \right| : x_i^{(k)} \in \mathbb{C},\ |x_i^{(k)}| \leq 1, \text{ for all } i \in \mathbb{N}, k \in \{1,2\} \right\} \geq \frac{1}{\sqrt{2}} \left(\sum_{i,j} |\beta_{i,j}|^{4/3} \right)^{3/4}. \tag{A12}$$

For $k > 2$, the above result resembles the Bohnenblust-Hille inequality, which states that, given $\{\beta_{i_1,\ldots,i_k} \in \mathbb{C}\}_{i_1,\ldots,i_k}$,

$$\sup \left\{ \left| \sum_{i_1,\ldots,i_k} \beta_{i_1,\ldots,i_k} x_{i_1}^{(1)} \cdots x_{i_k}^{(k)} \right| : x_{i_\kappa}^{(\kappa)} \in \mathbb{C},\ |x_{i_\kappa}^{(\kappa)}| \leq 1, \text{ for all } i_\kappa \in \mathbb{N}, \kappa \in [k] \right\} \geq D_k \left(\sum_{i_1,\ldots,i_k} |\beta_{i_1,\ldots,i_k}|^{2k/(k+1)} \right)^{(k+1)/(2k)} \tag{A13}$$

for some constant D_k that depends on k. For optimizing a general k-local Hamiltonian, the design of the randomized approximation algorithm is inspired by the original proof [56] of the Bohnenblust-Hille inequality from 1931, which is used to study the absolute convergence of the Dirichlet series.

b. Optimizing bounded-degree *k*-local Hamiltonians

Here, we consider a Hamiltonian given by a sum of k-qubit observables, where each qubit is acted on by at most d of the k-qubit observables. This is often referred to as a k-local Hamiltonian with a bounded degree d. We can combine Fact 2 and the main theorem to obtain the following corollary.

Corollary 6 (Optimizing a bounded-degree k-local Hamiltonian).—Given an n-qubit k-local Hamiltonian $H = \sum_{P:\,|P|\leq k} \alpha_P P$ with bounded degree d, $|\alpha_P| \leq 1$ for all P, and $k = \mathcal{O}(1)$, there is a randomized algorithm that runs in time $\mathcal{O}(nd)$ and produces either a random maximizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle} [\langle\psi| H |\psi\rangle] \geq \mathop{\mathbb{E}}_{|\phi\rangle:\text{ Haar}} [\langle\phi| H |\phi\rangle] + \frac{C}{\sqrt{d}} \sum_{P\neq I} |\alpha_P| \tag{A14}$$

or a random minimizing state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \leq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] - \frac{C}{\sqrt{d}} \sum_{P \neq I} |\alpha_P| \tag{A15}$$

for some constant C.

The task of optimizing bounded-degree k-local Hamiltonians has been considered in previous work [44].

Theorem 8 (Approximation guarantee from Ref. [44]).— Given an n-qubit 2-local Hamiltonian $H = \sum_{P:\,|P|\leq 2} \alpha_P P$ with bounded degree d and $|\alpha_P| \leq 1$ for all P, there is a polynomial-time randomized algorithm that produces a quantum circuit that generates a random maximizing state $|\psi\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \geq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] + \frac{C}{d}\left(\sum_{P \neq I} |\alpha_P|^2\right) \frac{\sum_{P \neq I} |\alpha_P|^2}{\sum_{P \neq I} \mathbb{1}[\alpha_P \neq 0]} \tag{A16}$$

as well as a random minimizing state $|\psi\rangle$ satisfying

$$\mathop{\mathbb{E}}_{|\psi\rangle}[\langle\psi| H |\psi\rangle] \leq \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] - \frac{C}{d}\left(\sum_{P \neq I} |\alpha_P|^2\right) \frac{\sum_{P \neq I} |\alpha_P|^2}{\sum_{P \neq I} \mathbb{1}[\alpha_P \neq 0]} \tag{A17}$$

for some constant C.

The result from Ref. [44] considers a single-step gradient descent using a shallow quantum circuit on an initial random product state. Because $\sum_{P\neq I} |\alpha_P|^2 \leq \sum_{P\neq I} \mathbb{1}[\alpha_P \neq 0]$ and $\sum_{P\neq I} |\alpha_P| \geq \sum_{P\neq I} |\alpha_P|^2$, our result in Corollary 6 improves either the maximization problem or the minimization problem over Theorem 8. For example, if we consider $\alpha_P = \Theta(1/d)$, which sets the total interaction strength on each qubit to be $\Theta(1)$, then the improvement over the Haar random state by our algorithm and that by the algorithm in Ref. [44] are given by

$$\Theta\left(\frac{1}{\sqrt{d}} \sum_{P\neq I} |\alpha_P|\right) = \Theta\left(\frac{n}{d^{1.5}}\right), \tag{A18}$$

$$\Theta\left(\frac{1}{\sqrt{d}}\left(\sum_{P \neq I} |\alpha_P|^2\right) \frac{\sum_{P \neq I} |\alpha_P|^2}{\sum_{P \neq I} \mathbb{1}[\alpha_P \neq 0]}\right) = \Theta\left(\frac{n}{d^{4.5}}\right). \tag{A19}$$

We can see that our algorithm gives a larger improvement for the scaling with degree d. As another example, consider a 1D n-qubit nearest-neighbor chain (hence $d = 2$), where $|\alpha_P| = 1$ for only a constant number of Pauli observables P and $|\alpha_P| = 1/\sqrt{n}$ for the rest of the Pauli observables. The improvement over the Haar random state by our algorithm and that by the algorithm in Ref. [44] are given by

$$\Theta\left(\frac{1}{\sqrt{d}} \sum_{P\neq I} |\alpha_P|\right) = \Theta(\sqrt{n}), \tag{A20}$$

$$\Theta\left(\frac{1}{\sqrt{d}}\left(\sum_{P \neq I} |\alpha_P|^2\right) \frac{\sum_{P \neq I} |\alpha_P|^2}{\sum_{P \neq I} \mathbb{1}[\alpha_P \neq 0]}\right) = \Theta\left(\frac{1}{n}\right). \tag{A21}$$

We can see that our algorithm gives a larger improvement for the scaling with the number n of qubits.

3. Description of the randomized approximation algorithm

There are a few steps in the proposed randomized algorithm. The first step is to choose the best slice of the k-local Hamiltonian by splitting the k-local Hamiltonian $H = \sum_{P:\,|P|\leq k} \alpha_P P$ as

$$H = \alpha_I I + \sum_{\kappa=1}^{k} H_\kappa, \qquad H_\kappa \triangleq \sum_{P:\,|P|=\kappa} \alpha_P P. \tag{A22}$$

We choose $\kappa^* \in \{1, \ldots, k\}$ to be the κ that maximizes $\sum_{P:\,|P|=\kappa} |\alpha_P|^r$, where $r = 2d_e/(d_e + 1)$. This step can be performed in time $\mathcal{O}(\mathsf{nnz}(H)k)$.

In the second step, the algorithm samples $(\kappa^* - 1)n$ Haar-random single-qubit pure states,

$$|\psi_{(s,j)}\rangle \in \mathbb{C}^2 \quad \text{for all } s \in \{1, \ldots, \kappa^* - 1\}, j \in \{1, \ldots, n\}. \tag{A23}$$

This step can be performed in time $\mathcal{O}(nk)$.

The third step is a local optimization on each qubit based on $|\psi_{(s,j)}\rangle$. For each qubit i and Pauli matrix $p \in \{X, Y, Z\}$, we define an $(n-1)$-qubit homogeneous $(\kappa^* - 1)$-local Hermitian operator

$$H_{\kappa^*,i,p} \triangleq \sum_{P \in \{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*, P_i = p} \alpha_P \left(\bigotimes_{j\neq i} P_j\right). \tag{A24}$$

For each qubit i and $p \in \{X, Y, Z\}$, the algorithm computes the real value, given as

$$\beta_{i,p} \triangleq \mathop{\mathbb{E}}_{\sigma\in\{\pm 1\}^{\kappa^*-1}}\left[\sigma_1 \cdots \sigma_{\kappa^*-1} \operatorname{tr}\left[H_{\kappa^*,i,p} \bigotimes_{j\neq i}\left[\frac{I}{2} + \frac{1}{\kappa^* - 1} \sum_{s=1}^{\kappa^*-1} \sigma_s \left(|\psi_{(s,j)}\rangle\langle\psi_{(s,j)}| - \frac{I}{2}\right)\right]\right]\right]. \tag{A25}$$

Then, for each qubit j, we consider a single-qubit local optimization

$$|\psi_{(\kappa^*,j)}\rangle \triangleq \underset{|\phi\rangle:\ \text{one-qubit state}}{\arg\max} \langle\phi| \left(\sum_{p\in\{X,Y,Z\}} \beta_{j,p} p\right) |\phi\rangle = \frac{I + n_X X + n_Y Y + n_Z Z}{2}, \tag{A26}$$

where $n_p = \beta_{j,p}/\sqrt{\sum_q \beta_{j,q}^2}$ for $p \in \{X, Y, Z\}$. After the optimization, the algorithm samples random numbers $\sigma_s \in \{\pm 1\}$ for all $s \in \{1, \ldots, \kappa^*\}$ to define a one-dimensional parameterized family of n-qubit product states,

$$\rho(t; |\psi_{(\cdot,\cdot)}\rangle, \sigma) \triangleq \bigotimes_{j=1}^{n} \left(\frac{I}{2} + \frac{t}{\kappa^*} \sum_{s=1}^{\kappa^*} \sigma_s \left(|\psi_{(s,j)}\rangle\langle\psi_{(s,j)}| - \frac{I}{2}\right)\right) \times \text{for all } t \in [-1, 1]. \tag{A27}$$

We denote this by $\rho(t)$ when $|\psi_{(\cdot,\cdot)}\rangle, \sigma$ are clear from the context. This concludes the third step. The third step can be performed in time $\mathcal{O}(\mathsf{nnz}(H)2^k)$.

The fourth step performs a polynomial optimization over the one-dimensional family

$$\max_{t\in[-1,1]} |\operatorname{tr}(H\rho(t; |\psi_{(\cdot,\cdot)}\rangle, \sigma)) - \alpha_I|. \tag{A28}$$

The function $f(t) = \operatorname{tr}(H\rho(t))$ is a polynomial of degree at most k. We can compute function $f(t)$ efficiently in time $\mathcal{O}(\mathsf{nnz}(H)k)$ as $\rho(t)$ is a product state. The optimization can thus be performed efficiently by sweeping through all possible values of t on a sufficiently fine grid. Let t^* be the optimal t.

The final step considers the sampling of a random pure state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$ from the distribution that corresponds to the mixed state $\rho(t^*; |\psi_{(\cdot,\cdot)}\rangle, \sigma)$. If $\operatorname{tr}(H\rho(t^*; |\psi_{(\cdot,\cdot)}\rangle, \sigma)) - \alpha_I > 0$ then the random product state $|\psi\rangle$ is a maximizing state satisfying Eq. (A3). Otherwise, the random product state $|\psi\rangle$ is a minimizing state satisfying Eq. (A4). This step can be performed in time $\mathcal{O}(n)$.

4. Proof of Theorem 5

The first step of the algorithm considers splitting the k-local Hamiltonian H into homogeneous κ-local Hamiltonians H_κ defined below. In particular, a homogeneous κ^*-local H_{κ^*} is chosen.

Definition 2 (Homogeneous k local).—A Hermitian operator H is homogeneous k local if $H = \sum_{P:\ |P|=k} \alpha_P P$.

The second step is a random sampling that generates a single-qubit pure state $|\psi_{(s,j)}\rangle$ for each qubit j and each copy $s \in \{1, \ldots, \kappa^* - 1\}$. The third step is the most important part of the proof. We devote Appendices A 4 a, A 4 b, and A 4 c to establishing the first inequality in (Corollary 9 below)

$$\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle} \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}} |\operatorname{tr}(H_{\kappa^*}\rho(t = 1; |\psi_{(\cdot,\cdot)}\rangle, \sigma))| \\
&\quad \geq \frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)} k^{k+1.5} \sqrt{6}^k} \left(\sum_{P:\ |P|=\kappa^*} |\alpha_P|^r\right)^{1/r} \\
&\quad \geq \frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)} k^{k+1.5+1/r} \sqrt{6}^k} \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}.
\end{aligned} \tag{A29}$$

The second inequality follows from $k\sum_{P:\ |P|=\kappa^*} |\alpha_P|^r \geq \sum_{\kappa=1}^{k} \sum_{P:\ |P|=\kappa} |\alpha_P|^r = \sum_{P\neq I} |\alpha_P|^r$. For the fourth step, the analysis of polynomial optimization given in Appendix A 4 d (Corollary 10) can be combined with the above inequality to obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle} \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}} |\operatorname{tr}(H\rho(t^*; |\psi_{(\cdot,\cdot)}\rangle, \sigma)) - \alpha_I| \\
&\quad \geq \frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)} k^{k+1.5+1/r} (\sqrt{6}(1+\sqrt{2}))^k} \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}.
\end{aligned} \tag{A30}$$

For the final step of the algorithm, using $\mathbb{E}_{|\psi\rangle}|\psi\rangle\langle\psi| = \rho(t^*; \rho_{(s,j)}, \sigma_s)$ and convexity, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle} \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}} \mathop{\mathbb{E}}_{|\psi\rangle} |\langle\psi| H |\psi\rangle - \alpha_I| \\
&\quad \geq \mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle} \mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}} \left|\operatorname{tr}\left(H \mathop{\mathbb{E}}_{|\psi\rangle} |\psi\rangle\langle\psi|\right) - \alpha_I\right| \\
&\quad \geq \frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)} k^{k+1.5+1/r} (\sqrt{6}+2\sqrt{3})^k} \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}.
\end{aligned} \tag{A31}$$

The theorem follows by noting that $\mathbb{E}_{|\phi\rangle:\ \text{Haar}}[\langle\phi| H |\phi\rangle] = \alpha_I$.

a. Polarization

We justify the definition of $\beta_{i,p}$ using polarization. Given an n-qubit homogeneous k-local observable $O = \sum_{P:\ |P|=k} \alpha_P P$, consider the following nk-qubit observable. First, we index the set $[nk]$ using ordered tuples (s, i), where $s \in [k]$ and $i \in [n]$. For every Pauli operator P on n qubits with $|P| = k$, suppose that it acts nontrivially on qubits $i_1 < \cdots < i_k$ via Pauli matrices $P_{i_1}, \ldots, P_{i_k}$. Then, for any permutation $\pi \in \mathcal{S}_k$, consider the nk-qubit observable $\mathsf{pol}_\pi(P)$ that acts on the $(\pi(s), i_s)$th qubit via P_{i_s} for all $s \in [k]$. Then define

$$\mathsf{pol}(P) := \frac{1}{k!}\sum_{\pi\in\mathcal{S}_k}\mathsf{pol}_\pi(P). \tag{A32}$$

We can extend $\mathsf{pol}(\cdot)$ linearly and define $\mathsf{pol}(O) \triangleq \sum_P \alpha_P \mathsf{pol}(P)$. We refer to $\mathsf{pol}(O)$ as the *polarization* of O. The squared Frobenius norm of O and $\mathsf{pol}(O)$ are related by

$$\mathrm{tr}(O^2) = k!\,\mathrm{tr}(\mathsf{pol}(O)^2). \tag{A33}$$

We prove the following operator analogue of the classical polarization identity.

Lemma 1 (Polarization identity).—For any nk-qubit product state $\rho = \bigotimes_{s\in[k]}[\bigotimes_{i\in[n]}\rho_{(s,i)}]$, any n-qubit homogeneous k-local observable O, and any $t \in \mathbb{R}$, we have the identity

$$t^k\,\mathrm{tr}(\mathsf{pol}(O)\rho) = \frac{k^k}{k!}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^k}\left[\sigma_1\cdots\sigma_k\,\mathrm{tr}\left(O\bigotimes_{i\in[n]}\left\{\frac{I}{2} + \frac{t}{k}\sum_{s=1}^{k}\sigma_s\left(\rho_{(s,i)}-\frac{I}{2}\right)\right\}\right)\right], \tag{A34}$$

where the expectation is with respect to the uniform measure on $\{\pm1\}^k$.

Proof.—Let $O = \sum_{P:\,|P|=k}\alpha_P P$. By the multinomial theorem, we can expand the right-hand side to get

$$\frac{t^k}{k!}\sum_{P:\,|P|=k}\alpha_P\mathop{\mathbb{E}}_{\sigma}\left[\sigma_1\cdots\sigma_k\sum_{0\le s_1,\ldots,s_n\le k}\mathrm{tr}\left(P\bigotimes_{i=1}^{n}\left\{\frac{I}{2}\mathbb{1}[s_i=0] + \sigma_{s_i}\left(\rho_{s_i,i}-\frac{I}{2}\right)\mathbb{1}[s_i>0]\right\}\right)\right]. \tag{A35}$$

For a given Pauli operator P, note that the only terms in the inner summation that are nonzero are given by $(s_1,\ldots,s_n)$ satisfying the condition that if $s_i > 0$ then P acts nontrivially on the ith qubit, because otherwise $\mathrm{tr}(\rho_{s_i,i} - I/2) = 0$ and the corresponding summand vanishes. Furthermore, for $(s_1,\ldots,s_n)$ satisfying this property, if $\{1,\ldots,k\}$ do not each appear exactly once then

$$\sigma_1\cdots\sigma_k\bigotimes_{i=1}^{n}\left\{\frac{I}{2}\mathbb{1}[s_i=0] + \sigma_{s_i}\left(\rho_{s_i,i}-\frac{I}{2}\right)\mathbb{1}[s_i>0]\right\}$$
$$= \sigma_1^{c_1}\cdots\sigma_k^{c_k}\bigotimes_{i=1}^{n}\left\{\frac{I}{2}\mathbb{1}[s_i=0] + \left(\rho_{s_i,i}-\frac{I}{2}\right)\mathbb{1}[s_i>0]\right\} \tag{A36}$$

for $0 \le c_1,\ldots,c_k \le k$ such that $c_s = 1$ for some $s \in [k]$. In this case, the expectation of this term with respect to σ vanishes. Altogether, we conclude that, for P that acts via $P_1,\ldots,P_k$ on qubits $1 \le i_1 < \cdots < i_k \le n$ and via identity elsewhere, the corresponding expectation over σ in Eq. (A35) is given by

$$\sum_{\pi\in\mathcal{S}_k}\mathrm{tr}\left(\bigotimes_{s=1}^{k}P_j\left(\rho_{\pi(s),i_s}-\frac{I}{2}\right)\right) = \sum_{\pi\in\mathcal{S}_k}\mathrm{tr}\left(\bigotimes_{s=1}^{k}P_s\rho_{\pi(s),i_s}\right)$$
$$= \sum_{\pi}\mathrm{tr}(\mathsf{pol}_\pi(P)\rho), \tag{A37}$$

from which the lemma follows. ■

Using the polarization identity, we can obtain the following corollary, which shows that $\beta_{i,p}$ is defined to be proportional to the expectation of the polarization $\mathsf{pol}(H_{\kappa^*,i,p})$ of the homogeneous κ^*-local observable $H_{\kappa^*,i,p}$ on the tensor product of $n(\kappa^*-1)$ single-qubit Haar-random states. We will later study the expectation value of the polarized observable on random product states.

Corollary 7.—From the definitions given in Appendix A 3, we have

$$\mathrm{tr}\left(\mathsf{pol}(H_{\kappa^*,i,p})\bigotimes_{s\in[\kappa^*-1],\,i\in[n]}|\psi_{(s,j)}\rangle\langle\psi_{(s,j)}|\right) = \frac{(\kappa^*-1)^{\kappa^*-1}}{(\kappa^*-1)!}\beta_{i,p}. \tag{A38}$$

Proof.—The claim follows from the polarization identity in Lemma 1 and the definition of $\beta_{i,p}$ in Eq. (A25). ■

b. Khintchine inequality for polarized observables

We recall the following basic result in high-dimensional probability.

Lemma 2 (Standard Khintchine inequality [79]).—Consider $\varepsilon_1,\ldots,\varepsilon_n$ to be independent and identically distributed random variables with $P(\varepsilon_i = \pm1) = 1/2$. For any $a_1,\ldots,a_n \in \mathbb{R}$, we have

$$\frac{1}{\sqrt{2}}\left(\sum_{i=1}^{n}a_i^2\right)^{1/2} \le \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left|\sum_{i=1}^{n}a_i\varepsilon_i\right| \le \left(\sum_{i=1}^{n}a_i^2\right)^{1/2}. \tag{A39}$$

We prove an analogue of the Khintchine inequality when we replace the random ±1 variables with random product states and replace $a_1,\ldots,a_n$ with a homogeneous 1-local observable.

Lemma 3 (Khintchine inequality for homogeneous 1-local observables).—Let $n \ge 1$. Consider $|\psi\rangle = \bigotimes_{i=1}^{n}|\psi_i\rangle$, where $|\psi_i\rangle$ is a single-qubit Haar-random pure state.

For any homogeneous 1-local n-qubit observable O,

$$\frac{1}{\sqrt{6}}\sqrt{\mathrm{tr}(O^2)/2^n} \leq \mathop{\mathbb{E}}_{|\psi\rangle}[|\langle\psi|\, O\,|\psi\rangle|] \leq \frac{1}{\sqrt{3}}\sqrt{\mathrm{tr}(O^2)/2^n}. \tag{A40}$$

Proof.—A homogeneous 1-local observable O is $\sum_{i=1}^{n}\sum_{j=1}^{3}\alpha_{ij}P_i^j$, where P_i^j is the Pauli matrix $\sigma_j \in \{X, Y, Z\}$ on the ith qubit. Given n single-qubit unitaries $U_1, \ldots, U_n$, we consider O under the rotated Pauli basis

$$O = \sum_{i=1}^{n}\sum_{j=1}^{3}\alpha_{ij}^{U}U_i^\dagger P_i^j U_i. \tag{A41}$$

Using the orthogonality of Pauli matrices, we have

$$\sqrt{\mathrm{tr}(O^2)/2^n} = \left(\sum_{i=1}^{n}\sum_{j=1}^{3}(a_{ij}^{U})^2\right)^{1/2} \tag{A42}$$

under any rotated Pauli basis. We utilize the rotated Pauli basis to establish the claimed results.

A single-qubit Haar-random pure state $|\psi_i\rangle$ can be sampled as follows. First, we sample a random single-qubit unitary U_i. Then, we consider $|\psi_i\rangle$ to be sampled uniformly from the set of eight pure states,

$$\Upsilon^{U_i} = \left\{\frac{I + (1/\sqrt{3})(s_i^X U_iXU_i^\dagger + s_i^Y U_iYU_i^\dagger + s_i^Z U_iZU_i^\dagger)}{2} \,\middle|\, \times\, s_i^X, s_i^Y, s_i^Z \in \{\pm 1\}\right\}. \tag{A43}$$

Using this sampling formulation and the rotated Pauli basis representation for O, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{|\psi\rangle}[|\langle\psi|\, O\,|\psi\rangle|] \\
&\quad = \mathop{\mathbb{E}}_{U_i}\mathop{\mathbb{E}}_{|\psi_i\rangle\sim\Upsilon^{U_i}}\left|\sum_{i=1}^{n}\sum_{j=1}^{3}\alpha_{ij}^{U}\,\mathrm{tr}(U_i^\dagger P_i^j U_i|\psi_i\rangle\langle\psi_i|)\right| \\
&\quad = \frac{1}{\sqrt{3}}\mathop{\mathbb{E}}_{U_i}\mathop{\mathbb{E}}_{s_i^X,s_i^Y,s_i^Z\sim\{\pm 1\}}\left|\sum_{i=1}^{n}\alpha_{i1}^{U}s_i^X + \alpha_{i2}^{U}s_i^Y + \alpha_{i3}^{U}s_i^Z\right|.
\end{aligned} \tag{A44}$$

Using the standard Khintchine inequality given in Lemma 2, we have

$$\begin{aligned}
&\frac{1}{\sqrt{2}}\left(\sum_{i=1}^{n}\sum_{j=1}^{3}(a_{ij}^{U})^2\right)^{1/2} \leq \mathop{\mathbb{E}}_{s_i^X,s_i^Y,s_i^Z\sim\{\pm 1\}} \\
&\quad \times \left|\sum_{i=1}^{n}\alpha_{i1}^{U}s_i^X + \alpha_{i2}^{U}s_i^Y + \alpha_{i3}^{U}s_i^Z\right| \leq \left(\sum_{i=1}^{n}\sum_{j=1}^{3}(a_{ij}^{U})^2\right)^{1/2}.
\end{aligned} \tag{A45}$$

Using Eq. (A42), we can obtain

$$\begin{aligned}
\frac{1}{\sqrt{6}}\mathop{\mathbb{E}}_{U_i}\sqrt{\mathrm{tr}(O^2)/2^n} &\leq \mathop{\mathbb{E}}_{|\psi\rangle}[|\langle\psi|\, O\,|\psi\rangle|] \\
&\leq \frac{1}{\sqrt{3}}\mathop{\mathbb{E}}_{U_i}\sqrt{\mathrm{tr}(O^2)/2^n},
\end{aligned} \tag{A46}$$

which implies the claimed result. ■

We prove the left half of the Khintchine inequality for polarized observables. The right half can be shown using a similar proof, but we are only going to use the left half stated below.

Lemma 4 (Khintchine inequality for polarized observables).—Let $n, k > 0$. Consider an nk-qubit observable $O = \mathsf{pol}(O')$, which is the polarization of an n-qubit homogeneous k-local observable O'. Consider $|\psi\rangle = \bigotimes_{s\in[k],i\in[n]}|\psi_{(s,i)}\rangle$, where $|\psi_{(s,i)}\rangle$ is a single-qubit Haar-random pure state. We have

$$\left(\frac{1}{\sqrt{6}}\right)^k\sqrt{\mathrm{tr}(O^2)/2^n} \leq \mathop{\mathbb{E}}_{|\psi\rangle}[|\langle\psi|\, O\,|\psi\rangle|]. \tag{A47}$$

Proof.—For $\ell \in [3n]$, define $P^{(\ell)}$ to be an n-qubit observable equal to the Pauli matrix $\sigma_{1+(\ell \bmod 3)} \in \{X, Y, Z\}$ acting on the $\lceil \ell/3 \rceil$th qubit. From the definition of polarization, we can represent O as

$$O = \sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_1)}\otimes\cdots\otimes P^{(\ell_k)}. \tag{A48}$$

For arbitrary coefficients $\alpha_{\ell_1,\ldots,\ell_k} \in \mathbb{R}$, we prove the following claim by induction on k:

$$\begin{aligned}
&\left(\frac{1}{\sqrt{6}}\right)^k\left(\sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}^2\right)^{1/2} \\
&\quad \leq \mathop{\mathbb{E}}_{|\psi\rangle}\left[\left|\langle\psi|\sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_1)}\otimes\cdots\otimes P^{(\ell_k)}\,|\psi\rangle\right|\right].
\end{aligned} \tag{A49}$$

It is not hard to see that the left-hand side of Eq. (A49) is $\left(1/\sqrt{6}\right)^k\sqrt{\mathrm{tr}(O^2)/2^n}$ and the right-hand side of Eq. (A49) is $\mathbb{E}_{|\psi\rangle}[|\langle\psi|\, O\,|\psi\rangle|]$. Hence, the lemma follows from Eq. (A49).

We now prove the base case and the inductive step. The base case of $k = 1$ follows from the Khintchine inequality for homogeneous 1-local observables given in Lemma 3. Assume by the induction hypothesis that the claim holds for $k - 1$. Denoting by $|\psi^{(k)}\rangle$ the product of n Haar-random single-qubit states, we can then apply the Khintchine inequality for homogeneous 1-local observables (Lemma 3) to obtain

$$\left(\frac{1}{\sqrt{6}}\right)^k\left(\sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha^2_{\ell_1,\ldots,\ell_k}\right)^{1/2}=\left(\sum_{\ell_1,\ldots,\ell_{k-1}\in[3n]}\left(\left(\sum_{\ell_k\in[3n]}\alpha^2_{\ell_1,\ldots,\ell_k}\right)^{1/2}\right)^2\right)^{1/2}$$
$$\leq\left(\sum_{\ell_1,\ldots,\ell_{k-1}\in[3n]}\left(\mathop{\mathbb{E}}_{|\psi^{(k)}\rangle}\left|\langle\psi^{(k)}|\sum_{\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_k)}|\psi^{(k)}\rangle\right|\right)^2\right)^{1/2}. \tag{A50}$$

We can then apply Minkowski's integral inequality to the upper bound above and yield

$$\left(\frac{1}{\sqrt{6}}\right)^k\left(\sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha^2_{\ell_1,\ldots,\ell_k}\right)^{1/2}\leq\mathop{\mathbb{E}}_{|\psi^{(k)}\rangle}\left(\sum_{\ell_1,\ldots,\ell_{k-1}\in[3n]}\left(\langle\psi^{(k)}|\sum_{\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_k)}|\psi^{(k)}\rangle\right)^2\right)^{1/2}$$
$$\leq\mathop{\mathbb{E}}_{|\psi^{(k)}\rangle}\mathop{\mathbb{E}}_{|\psi^{(1,\ldots,k-1)}\rangle}\Bigg|\langle\psi^{(1,\ldots,k-1)}|\langle\psi^{(k)}|\sum_{\ell_1,\ldots,\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_1)}\otimes\cdots$$
$$\otimes P^{(\ell_k)}|\psi^{(1,\ldots,k-1)}\rangle|\psi^{(k)}\rangle\Bigg|. \tag{A51}$$

The last inequality considers $\langle\psi^{(k)}|\sum_{\ell_k\in[3n]}\alpha_{\ell_1,\ldots,\ell_k}P^{(\ell_k)}|\psi^{(k)}\rangle$ to be a scalar indexed by $\ell_1,\ldots,\ell_{k-1}$ and uses the induction hypothesis. We have thus established the induction step. The claim in Eq. (A49) follows. ■

The Khintchine inequality for polarized observables allows us to show that the average magnitude of $\mathsf{pol}(H_{\kappa^*,i,p})$ for the tensor product of single-qubit Haar-random states is at least as large as the Frobenius norm of $H_{\kappa^*,i,p}$ up to a constant depending on κ^*. Using the definitions from the design of the approximate optimization algorithm, we can obtain the following corollary.

Corollary 8.—From the definitions given in Appendix A 3, we have

$$\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\left|\mathrm{tr}\left(\mathsf{pol}(H_{\kappa^*,i,p})\bigotimes_{s\in[\kappa^*-1],\,i\in[n]}|\psi_{(s,j)}\rangle\langle\psi_{(s,j)}|\right)\right|$$
$$\geq\left(\frac{1}{\sqrt{6}}\right)^{\kappa^*-1}\sqrt{\frac{\mathrm{tr}(H^2_{\kappa^*,i,p})}{2^n(\kappa^*-1)!}}. \tag{A52}$$

Proof.—The claim follows immediately from Lemma 4 and Eq. (A33). ■

c. Characterization of the locally optimized random state

Recall that $\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma)$ is created by sampling random product states and performing local single-qubit optimizations. The locally optimized random state satisfies the following inequality.

Lemma 5 (Characterization of $\rho(t)$ for $t=1$).—From the definitions given in Appendix A 3, we have

$$\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}|\mathrm{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|$$
$$\geq\frac{\sqrt{2(\kappa^*!)}}{(\kappa^*)^{\kappa^*+1.5}\sqrt{6}^{\kappa^*}}\sum_{i\in[n],\,p\in\{X,Y,Z\}}$$
$$\times\sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*,P_i=p}\alpha_P^2}. \tag{A53}$$

Proof.—From the polarization identity given in Lemma 1, we have

$$\frac{(\kappa^*)^{\kappa^*}}{\kappa^*!}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}[\sigma_1\cdots\sigma_{\kappa^*}\,\mathrm{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))]$$
$$=\mathrm{tr}\left(\mathsf{pol}(H_{\kappa^*})\bigotimes_{s\in[\kappa^*],\,j\in[n]}|\psi_{(s,j)}\rangle\langle\psi_{(s,j)}|\right). \tag{A54}$$

Next, using the definition of $H_{\kappa^*,i,p}$ in Eq. (A24), we have

$$\mathsf{pol}(H_{\kappa^*})=\left(\frac{1}{\kappa^*}\right)^2\sum_{i\in[n]}\sum_{p\in\{X,Y,Z\}}\mathsf{pol}(H_{\kappa^*,i,p})$$
$$\otimes(I^{\otimes i-1}\otimes p\otimes I^{\otimes n-i}). \tag{A55}$$

We can see this by considering the case when H_{κ^*} is a single Pauli observable $P\in\{I,X,Y,Z\}^{\otimes n}$ with $|P|=\kappa^*$, and then extending linearly to any homogeneous κ^*-local

Hamiltonian H_{κ^*}. Equations (A54) and (A55) give

$$\begin{aligned}\frac{(\kappa^*)^{\kappa^*}}{\kappa^*!}&\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}[\sigma_1\cdots\sigma_{\kappa^*}\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))]\\ &=\frac{1}{(\kappa^*)^2}\sum_{i\in[n],p\in\{X,Y,Z\}}\langle\psi_{(\kappa^*,i)}|\,p\,|\psi_{(\kappa^*,i)}\rangle\\ &\quad\times\operatorname{tr}\left(\mathsf{pol}(H_{\kappa^*,i,p})\bigotimes_{s\in[\kappa^*-1],j\in[n]}|\psi_{(s,j)}\rangle\!\langle\psi_{(s,j)}|\right).\end{aligned}\tag{A56}$$

From Corollary 7, we can rewrite the right-hand side as

$$\frac{1}{(\kappa^*)^2}\frac{(\kappa^*-1)^{\kappa^*-1}}{(\kappa^*-1)!}\sum_{i\in[n]}\langle\psi_{(\kappa^*,i)}|\left(\sum_{p\in\{X,Y,Z\}}\beta_{i,p}p\right)|\psi_{(\kappa^*,i)}\rangle\,.\tag{A57}$$

From the local optimization of $|\psi_{(\kappa^*,i)}\rangle$ given in Eq. (A26), we have, for every $i\in[n]$,

$$\begin{aligned}\langle\psi_{(\kappa^*,i)}|\left(\sum_{p\in\{X,Y,Z\}}\beta_{i,p}p\right)|\psi_{(\kappa^*,i)}\rangle&=\sqrt{\sum_{p\in\{X,Y,Z\}}\beta_{i,p}^2}\\ &\geq\frac{1}{\sqrt{3}}\sum_{p\in\{X,Y,Z\}}|\beta_{i,p}|.\end{aligned}\tag{A58}$$

Using Corollary 7 yields the lower bound

$$\begin{aligned}\frac{(\kappa^*)^{\kappa^*}}{\kappa^*!}&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}[\sigma_1\cdots\sigma_{\kappa^*}\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))]\\ &\geq\frac{1}{\sqrt{3}(\kappa^*)^2}\sum_{i\in[n],p\in\{X,Y,Z\}}\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\\ &\quad\times\left|\operatorname{tr}\left(\mathsf{pol}(H_{\kappa^*,i,p})\bigotimes_{s\in[\kappa^*-1],j\in[n]}|\psi_{(s,j)}\rangle\!\langle\psi_{(s,j)}|\right)\right|.\end{aligned}\tag{A59}$$

From Corollary 8, we can further obtain

$$\begin{aligned}\frac{(\kappa^*)^{\kappa^*}}{\kappa^*!}&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}[\sigma_1\cdots\sigma_{\kappa^*}\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))]\\ &\geq\frac{1}{\sqrt{3}(\kappa^*)^2}\sum_{i\in[n],p\in\{X,Y,Z\}}\left(\frac{1}{\sqrt{6}}\right)^{\kappa^*-1}\sqrt{\frac{\operatorname{tr}(H_{\kappa^*,i,p}^2)}{2^n(\kappa^*-1)!}}.\end{aligned}\tag{A60}$$

The definition of $H_{\kappa^*,i,p}$, the above inequality, and the inequality

$$\begin{aligned}\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}&\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}|\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|\\ &\geq\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\mathop{\mathbb{E}}_{\sigma\in\{\pm1\}^{\kappa^*}}[\sigma_1\cdots\sigma_{\kappa^*}\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))]\end{aligned}\tag{A61}$$

can be used to establish the claim. ■

Given the expansion property, we are going to use the following implication, which considers an arbitrary ordering π of the n qubits. The inequality allows us to control the growth for the number of Pauli observables that act on qubits before the ith qubit under ordering π. The precise statement is given below.

Lemma 6 (A characterization of expansion).—Suppose that there is an n-qubit Hamiltonian $H=\sum_P\alpha_PP$ with expansion coefficient c_e and expansion dimension d_e. Consider any permutation $\pi\in S_n$ over n qubits. For any $i\in[n]$,

$$\begin{aligned}&\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\mathbb{1}[\alpha_P\neq0]\mathbb{1}[P_{\pi(i)}\neq I]\mathbb{1}[P_{\pi(j)}\\ &\quad=I\text{ for all }j>i]\leq c_ei^{d_e-1}.\end{aligned}\tag{A62}$$

Proof.—Consider a permutation $\pi\in S_n$ over n qubits and an $i\in[n]$. We separately consider two cases: (1) $i<d_e$ and (2) $i\geq d_e$. For the first case, let $\Upsilon=\{\pi(1),\ldots,\pi(d_e)\}$; we have

$$\begin{aligned}&\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\mathbb{1}[\alpha_P\neq0]\mathbb{1}[P_{\pi(i)}\neq I]\mathbb{1}[P_{\pi(j)}=I\text{ for all }j>i]\\ &\quad\leq\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\mathbb{1}[\alpha_P\neq0\text{ and }(\mathsf{dom}(P)\subseteq\Upsilon)]\\ &\quad\leq c_e.\end{aligned}\tag{A63}$$

The second inequality follows from the definition of the expansion coefficient c_e. For the second case, we consider all subsets $\Upsilon\subseteq\pi([i])\triangleq\{\pi(1),\pi(2),\ldots,\pi(i)\}$ with $|\Upsilon|=d_e-1$ and $\pi(i)\in\Upsilon$; we have

$$\begin{aligned}&\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\mathbb{1}[\alpha_P\neq0]\mathbb{1}[P_{\pi(i)}\neq I]\mathbb{1}[P_{\pi(j)}=I\text{ for all }j>i]\\ &\quad\leq\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\sum_{\substack{\Upsilon\subseteq\pi([i]),\\|\Upsilon|=d_e,\pi(i)\in\Upsilon}}\mathbb{1}[\alpha_P\neq0\text{ and }(\mathsf{dom}(P)\\ &\quad\subseteq\Upsilon\text{ or }\Upsilon\subseteq\mathsf{dom}(P))]\\ &\quad\leq\sum_{\substack{\Upsilon\subseteq\pi([i]),\\|\Upsilon|=d_e,\pi(i)\in\Upsilon}}c_e\leq c_e(i-1)^{d_e-1}\\ &\quad\leq c_ei^{d_e-1}.\end{aligned}\tag{A64}$$

The second inequality again follows from the definition of c_e. ■

Using the above implication of the expansion property, we can obtain the following inequality relating two norms. Basically, we can use the limit on the growth of the number of Pauli observables to turn the sum of the ℓ_2 norm into an ℓ_r norm, where r depends on the expansion dimension d_e.

Lemma 7 (Norm inequality using the expansion property).—Consider an n-qubit Hamiltonian $H = \sum_P \alpha_P P$ with an expansion coefficient c_e and expansion dimension d_e. Let $r = 2d_e/(d_e + 1)$. For any $\kappa^* \geq 1$, we have

$$\sum_{i\in[n]} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_i\neq I} \alpha_P^2 \right)^{1/2} \geq \frac{1}{c_e^{1/(2d_e)}} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*} |\alpha_P|^r \right)^{1/r}. \tag{A65}$$

Proof.—We begin by considering a permutation π over n qubits such that

$$\sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2} \leq \sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(j)}\neq I} \alpha_P^2} \quad \text{for all } i < j \in [n]. \tag{A66}$$

Permutation π can be obtained by sorting the n qubits. The above ensures that, for all $i \in [n]$,

$$i \sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2} \leq \sum_{j\in[n]} \sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(j)}\neq I} \alpha_P^2}. \tag{A67}$$

By going through the n qubits based on permutation π, we have the identity

$$\sum_{P:\, |P|=\kappa^*} |\alpha_P|^r = \sum_{i=1}^{n} \sum_{p\in\{X,Y,Z\}} \sum_{\substack{P\in\{I,X,Y,Z\}^{\otimes n}: \\ |P|=\kappa^*, P_{\pi(i)}=p}} |\alpha_P|^r \mathbb{1}[\alpha_P \neq 0] \mathbb{1}[P_{\pi(j)} = I \text{ for all } j > i]. \tag{A68}$$

Holder’s inequality and $1/(d_e + 1) = 1 - r/2$ allows us to obtain the following upper bound on $\sum_{P:\, |P|=\kappa^*} |\alpha_P|^r$:

$$\sum_{i=1}^{n} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2 \right)^{r/2} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*} \mathbb{1}[\alpha_P \neq 0] \mathbb{1}[P_{\pi(i)} \neq I] \mathbb{1}[P_{\pi(j)} = I \text{ for all } j > i] \right)^{1/(d_e+1)}. \tag{A69}$$

We can then use Lemma 6 to obtain

$$\sum_{P:\, |P|=\kappa^*} |\alpha_P|^r \leq \sum_{i=1}^{n} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2 \right)^{r/2} (c_e i^{d_e-1})^{1/(d_e+1)}. \tag{A70}$$

Using $r - 1 = (d_e - 1)/(d_e + 1) \geq 0$, we have

$$\sum_{P:\, |P|=\kappa^*} |\alpha_P|^r \leq c_e^{1/(d_e+1)} \sum_{i=1}^{n} \left(i \sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2} \right)^{r-1} \sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_{\pi(i)}\neq I} \alpha_P^2}. \tag{A71}$$

The choice of π ensures Eq. (A67), which gives rise to

$$\sum_{P:\, |P|=\kappa^*} |\alpha_P|^r \leq c_e^{1/(d_e+1)} \left(\sum_{i\in[n]} \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\, |P|=\kappa^*, P_i\neq I} \alpha_P^2 \right)^{1/2} \right)^{r}. \tag{A72}$$

The claim follows from $1/(r(d_e + 1)) = 1/(2d_e)$. ■

Together, we can obtain the ℓ_r-norm lower bound for the expectation value of the homogeneous κ^*-local Hamiltonian H_{κ^*} on the constructed product state $\rho(1; |\psi_{(\cdot,\cdot)}\rangle, \sigma)$.

Corollary 9.—From the definitions given in Appendix A 3, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\ \mathop{\mathbb{E}}_{\sigma\in\{\pm 1\}^{\kappa^*}} |\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|\\ &\geq \frac{\sqrt{2(\kappa^*!)}}{c_e^{1/(2d_e)}(\kappa^*)^{\kappa^*+1.5}\sqrt{6}^{\kappa^*}}\left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*}|\alpha_P|^r\right)^{1/r}\\ &\geq \frac{\sqrt{2(k!)}}{c_e^{1/(2d_e)}k^{k+1.5}\sqrt{6}^{k}}\left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*}|\alpha_P|^r\right)^{1/r}.\end{aligned} \tag{A73}$$

Proof.—From Lemma 5, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{|\psi_{(\cdot,\cdot)}\rangle}\ \mathop{\mathbb{E}}_{\sigma\in\{\pm 1\}^{\kappa^*}} |\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|\\ &\quad\geq \frac{\sqrt{2(\kappa^*!)}}{(\kappa^*)^{\kappa^*+1.5}\sqrt{6}^{\kappa^*}}\sum_{i\in[n],p\in\{X,Y,Z\}}\\ &\quad\times\sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*,P_i=p}\alpha_P^2}.\end{aligned} \tag{A74}$$

By the elementary inequality $\sqrt{x}+\sqrt{y}+\sqrt{z}\geq\sqrt{x+y+z}$ for non-negative x,y,z,

$$\begin{aligned}&\sum_{i\in[n],p\in\{X,Y,Z\}}\sqrt{\sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*,P_i=p}\alpha_P^2}\\ &\quad\geq\sum_{i\in[n]}\sqrt{\sum_{p\in\{X,Y,Z\}}\ \sum_{P\in\{I,X,Y,Z\}^{\otimes n}:\,|P|=\kappa^*,P_i=p}\alpha_P^2}.\end{aligned} \tag{A75}$$

Combining this result with Lemma 7 and the fact that $k\geq\kappa^*$ yields the stated result. ■

d. Homogeneous to inhomogeneous through polynomial optimization

We need the following basic result from real analysis.

Lemma 8 (Markov brothers' inequality; see, e.g., [80, p. 248).—]

For any real polynomial $p(t)=\sum_{\kappa=1}^{k}a_\kappa x^\kappa$,

$$|a_\kappa|\leq(1+\sqrt{2})^k\sup_{|t|\leq 1}|p(t)| \tag{A76}$$

for all $1\leq\kappa\leq k$.

Using the Markov brothers' inequality, we can show that performing the one-dimensional polynomial optimization over t achieves a good advantage over $\alpha_I=\mathbb{E}_{|\psi\rangle:\,\text{Haar}}\langle\psi|H|\psi\rangle$.

Corollary 10.—From the definitions given in Appendix A 3, we have

$$\begin{aligned}&|\operatorname{tr}(H\rho(t^*;|\psi_{(\cdot,\cdot)}\rangle,\sigma))-\alpha_I|\\ &\quad\geq\frac{1}{(1+\sqrt{2})^k}|\operatorname{tr}(H_{\kappa^*}\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|.\end{aligned} \tag{A77}$$

Proof.—Recall that $H=\alpha_I I+\sum_{\kappa=1}^{k}H_\kappa$ from Eq. (A22). We can use the polarization identity given in Lemma 1 to see that the function $f(t)=\operatorname{tr}(H\rho(t;|\psi_{(\cdot,\cdot)}\rangle,\sigma))$ is a polynomial:

$$\operatorname{tr}(H\rho(t;|\psi_{(\cdot,\cdot)}\rangle,\sigma))=\alpha_I+\sum_{\kappa=1}^{k}\operatorname{tr}(H_\kappa\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))t^\kappa. \tag{A78}$$

Recall that t^* is chosen based on the optimization

$$\max_{t\in[-1,1]}|\operatorname{tr}(H\rho(t;|\psi_{(\cdot,\cdot)}\rangle,\sigma))-\alpha_I|. \tag{A79}$$

By considering Lemma 8 with $a_\kappa=\operatorname{tr}(H_\kappa\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))$, we have

$$\begin{aligned}&(1+\sqrt{2})^k|\operatorname{tr}(H\rho(t^*;|\psi_{(\cdot,\cdot)}\rangle,\sigma))-\alpha_I|\\ &\quad\geq|\operatorname{tr}(H_\kappa\rho(1;|\psi_{(\cdot,\cdot)}\rangle,\sigma))|.\end{aligned} \tag{A80}$$

This concludes the proof of this corollary. ■

APPENDIX B: NORM INEQUALITIES FROM THE APPROXIMATE OPTIMIZATION ALGORITHM

The approximate optimization algorithm described in the previous section is not used directly in the ML algorithm, but used to derive norm inequalities, i.e., inequalities relating different norms over Hermitian operators. An important norm that we use in the ML algorithms is the Pauli-p norm defined below. The Pauli-p norm is equivalent to the vector-p norm on the Pauli coefficient of an observable H.

Definition 3 (Pauli-p norm).—Given $H=\sum_{P\in\{I,X,Y,Z\}^{\otimes n}}\alpha_P P$ and $p\geq 1$, the Pauli-p norm of H is

$$\|H\|_{\text{Pauli},p}=\left(\sum_P|\alpha_P|^p\right)^{1/p}. \tag{B1}$$

Recall that the spectral norm $\|H\|=\max_{|\psi\rangle}|\langle\psi|H|\psi\rangle|=\max_\rho|\operatorname{tr}(H\rho)|$. In this section, we use the approximate optimization algorithm to derive several norm inequalities relating the Pauli-p norm $\|\cdot\|_{\text{Pauli},p}$ to the spectral norm $\|\cdot\|$ for common classes of observables.

We begin with a well-known fact that equates the Frobenius norm and the Pauli-2 norm. This proposition follows directly from the orthonormality of the Pauli observables $\{I, X, Y, Z\}^{\otimes n}$.

Proposition 1 (Frobenius norm).—Given any n-qubit Hermitian operator H, we have

$$\frac{1}{\sqrt{2^n}}\|H\|_F = \|H\|_{\text{Pauli},2} \leq \|H\|. \tag{B2}$$

Proof.—Let n be the number of qubits that H acts on, and let $\lambda_1, \ldots, \lambda_{2^n}$ be the eigenvalues of O. From the fact that $\mathrm{tr}(PQ) = 2^n \delta_{P=Q}$, we have

$$\|H\|_F^2 = \mathrm{tr}(H^2) = \sum_P |\alpha_P|^2 2^n = 2^n \|H\|_{\text{Pauli},2}^2. \tag{B3}$$

Since $\|H\|_F^2 = \sum_{i=1}^{2^n} |\lambda_i|^2 \leq 2^n \max_i |\lambda_i|^2 = 2^n \|H\|_\infty^2$, we have $\sum_P |\alpha_P|^2 = \|H\|_F^2/2^n \leq \|H\|_\infty^2$. ■

We now utilize Theorem 5 to obtain the following useful norm inequality.

Theorem 9 (Norm inequality from Theorem 5).—Consider an n-qubit k-local Hamiltonian H with expansion coefficient c_e and dimension d_e. Let $r = 2d_e/(d_e+1) \in [1, 2)$. We have

$$\tfrac{1}{3} C(c_e, d_e, k)\|H\|_{\text{Pauli},r} \leq \|H\|, \tag{B4}$$

where $C(c_e, d_e, k) = \sqrt{2(k!)}/[c_e^{1/(2d_e)} k^{k+1.5+1/r}(\sqrt{6} + 2\sqrt{3})^k]$ is the same as in Theorem 5.

Proof.—Consider the Pauli representation $H = \sum_{P:\,|P|\leq k} \alpha_P P$. If we consider $\rho = I/2^n$ then we have

$$\|H\| \geq |\,\mathrm{tr}(H)/2^n| \geq \left|\mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}} [\langle\phi| H |\phi\rangle]\right| = |\alpha_I|. \tag{B5}$$

If we consider the random product state $|\psi\rangle$ from Theorem 5 then we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{|\psi\rangle}\left|\langle\psi| H |\psi\rangle - \mathop{\mathbb{E}}_{|\phi\rangle:\,\text{Haar}} [\langle\phi| H |\phi\rangle]\right| \\ &\quad \geq C(c_e, d_e, k)\left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}.\end{aligned} \tag{B6}$$

Using $\mathbb{E}_{|\phi\rangle:\,\text{Haar}}[\langle\phi| H |\phi\rangle] = \alpha_I$ and $\mathbb{E}_{|\psi\rangle}|\,\langle\psi| H |\psi\rangle - \alpha_I| \leq \mathbb{E}_{|\psi\rangle}|\,\langle\psi| H |\psi\rangle\,| + |\alpha_I|$, we have

$$\|H\| \geq \mathop{\mathbb{E}}_{|\psi\rangle} |\langle\psi| H |\psi\rangle| \geq C(c_e, d_e, k)\left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r} - |\alpha_I|. \tag{B7}$$

Next, we utilize the inequality

$$\max(x_1, cx_2 - x_1) \geq \frac{c}{c+2}(x_1 + x_2) \quad \text{for all } x_1, x_2, c \geq 0, \tag{B8}$$

which can be shown by considering the two cases $x_1 \geq (c/2)x_2$ and $x_1 < (c/2)x_2$, as well as the lower bounds on $\|H\|$ to show that

$$\begin{aligned}\|H\| &\geq \frac{C(c_e, d_e, k)}{C(c_e, d_e, k) + 2}\left(|\alpha_I| + \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}\right) \\ &\geq \frac{C(c_e, d_e, k)}{3}\left(|\alpha_I| + \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r}\right).\end{aligned} \tag{B9}$$

The second inequality uses $k, c_e, d_e \geq 1$, which implies that $C(c_e, d_e, k) \in [0, 1]$. Finally, the inequality

$$|\alpha_I| + \left(\sum_{P\neq I} |\alpha_P|^r\right)^{1/r} \geq \left(\sum_{P} |\alpha_P|^r\right)^{1/r} \tag{B10}$$

can be used to establish the claim. ■

Using Facts 1 and 2 that characterize the expansion property for general k-local Hamiltonians and bounded-degree k-local Hamiltonians (i.e., each qubit is acted on by at most d of the k-qubit observables), we can establish the following corollaries.

Corollary 11 (Norm inequality for a k-local Hamiltonian).—Given an n-qubit k-local Hamiltonian H, we have

$$\tfrac{1}{3} C(k)\|H\|_{\text{Pauli},2k/(k+1)} \leq \|H\|, \tag{B11}$$

where $C(k) = \sqrt{2(k!)}/[2k^{k+1.5+(k+1)/(2k)}(\sqrt{6} + 2\sqrt{3})^k]$ is the same as in Corollary 5.

Corollary 12 (Norm inequality for a bounded-degree Hamiltonian).—Given an n-qubit k-local Hamiltonian H with bounded degree d, we have

$$\tfrac{1}{3} C(k, d)\|H\|_{\text{Pauli},1} \leq \|H\|, \tag{B12}$$

where $C(k, d) = \sqrt{2(k!)}/[\sqrt{d}k^{k+2.5}(2\sqrt{6} + 4\sqrt{3})^k]$.

APPENDIX C: SAMPLE-OPTIMAL ALGORITHMS FOR PREDICTING BOUNDED-DEGREE OBSERVABLES

In this appendix, we consider one of the most basic learning problems in quantum information theory: predicting properties of an unknown n-qubit state ρ. This has been studied extensively in the literature on shadow tomography [57,58] and classical shadows [46].

1. Review of classical shadow formalism

We recall the following definition and theorem from classical shadow tomography [46] based on randomized Pauli measurements. Each randomized Pauli measurement is performed on a single copy of ρ and measures each qubit of ρ in a random Pauli basis (X, Y, or Z).

Definition 4 (Shadow norm from randomized Pauli measurements).—Consider an n-qubit observable O. Let $\mathcal{U}$ be the distribution over the tensor product of n single-qubit random Clifford unitaries, and let $\mathcal{M}_P^{-1} = \bigotimes_{i=1}^n \mathcal{M}_1^{-1}$ with $\mathcal{M}_1^{-1}(A) = 3A - \mathrm{tr}(A)I$. The shadow norm of O is defined as

$$\|O\|_{\text{shadow}} = \max_{\sigma:\ \text{state}} \Bigg(\mathop{\mathbb{E}}_{U\sim\mathcal{U}} \sum_{b\in\{0,1\}^n} \langle b|\, U\sigma U^\dagger \\ |b\rangle\, \langle b|\, U\mathcal{M}_P^{-1}(O)U^\dagger\, |b\rangle^2 \Bigg)^{1/2}. \quad \text{(C1)}$$

Theorem 10 (Classical shadow tomography using randomized Pauli measurements [46]).—Given an unknown n-qubit state ρ and M observables $O_1, \ldots, O_M$ with $B_{\text{shadow}} = \max_{i\in[M]}\|O_i\|_{\text{shadow}}$, after N randomized Pauli measurements on copies of ρ satisfying

$$N = \mathcal{O}\left(\frac{\log(M)B_{\text{shadow}}^2}{\epsilon^2}\right), \quad \text{(C2)}$$

we can estimate $\mathrm{tr}(O_i\rho)$ to ϵ error for all $i \in [M]$ with high probability.

We can see that the sample complexity for predicting many properties of an unknown quantum state ρ depends on the shadow norm $\|\cdot\|_{\text{shadow}}$. The larger $\|\cdot\|_{\text{shadow}}$ is, the more experiments are needed to estimate the properties of ρ accurately. From the original classical shadow paper [46], we can obtain the following shadow-norm bounds for Pauli observables and for few-body observables.

Lemma 9 (Shadow norm for Pauli observables [46]).—For any $P \in \{I, X, Y, Z\}^{\otimes n}$, we have

$$\|P\|_{\text{shadow}} = 3^{|P|/2}. \quad \text{(C3)}$$

Lemma 10 (Shadow norm for few-body observables [46]).—For any observable O that acts nontrivially on at most k qubits, we have

$$\|O\|_{\text{shadow}} \le 2^k\|O\|. \quad \text{(C4)}$$

Combining the above lemmas and Theorem 10, we can see that Pauli observables and few-body observables can both be predicted efficiently under a very small number of randomized Pauli measurements.

2. Upper bound for predicting bounded-degree observables

Consider an n-qubit observable O given as a sum of k-qubit observables $O = \sum_j O_j$, where each qubit is acted on by at most d of these k-qubit observables O_j. We focus on $k = \mathcal{O}(1)$ and $d = \mathcal{O}(1)$, and refer to such an observable as a bounded-degree observable. These bounded-degree observables arise frequently in quantum many-body physics and quantum information. For example, the Hamiltonian in a quantum spin system can often be described by a geometrically local Hamiltonian, which is an instance of bounded-degree observables. For these observables, the shadow norm is related to the Pauli-1 norm of the observable:

$$\|O\|_{\text{shadow}} \le \sum_{P:\ |P|\le k} |\alpha_P|\,\|P\|_{\text{shadow}} \le 3^{k/2} \\ \sum_{P:\ |P|\le k} |\alpha_P| = 3^{k/2}\|O\|_{\text{Pauli},1}. \quad \text{(C5)}$$

If we consider the norm inequality between the ℓ_1 norm and ℓ_2 norm and use the standard result relating the Frobenius norm and spectral norm (Proposition 1), we would obtain the following upper bound on the shadow norm:

$$\|O\|_{\text{shadow}} \le 3^{k/2}\|O\|_{\text{Pauli},1} \le (2\sqrt{3})^k\sqrt{nd}\|O\|_{\text{Pauli},2} \\ = \mathcal{O}(\sqrt{n}\|O\|). \quad \text{(C6)}$$

Using Theorem 10, this shadow-norm bound gives rise to a number of measurements scaling as

$$N = \mathcal{O}\left(\frac{n\log(M)B_\infty^2}{\epsilon^2}\right), \quad \text{(C7)}$$

where $B_\infty = \max_{i\in[M]}\|O_i\|_\infty$ is an upper bound on the spectral norm $\|\cdot\|$. Because of the linear dependence on the number n of qubits in the unknown quantum state, this scaling is not ideal. Furthermore, we will later show that this scaling is actually far from optimal.

To improve the sample complexity, we use the improved approximate optimization algorithm presented in Appendix A, and the corresponding norm inequality presented in Appendix B. Using the norm inequality relating the Pauli-1 norm and the spectral norm (Corollary 12), we can obtain the following shadow-norm bound.

Lemma 11 (Shadow norm for bounded-degree observables).—Given $k, d = \mathcal{O}(1)$ and an n-qubit observable O that is a sum of k-qubit observables, where each qubit is acted on by at most d of these k-qubit observables,

$$\|O\|_{\text{shadow}} \le C\|O\| \quad \text{(C8)}$$

for some constant $C > 0$.

Combining the above lemma with Theorem 10 allows us to establish the following theorem. Compared to Eq. (C7), the following theorem uses n times fewer measurements.

Theorem 11 (Classical shadow tomography for bounded-degree observables).—Consider an unknown n-qubit state ρ and M observables $O_1, \ldots, O_M$ with $B_\infty = \max_i\|O_i\|_\infty$. Suppose that each observable O_i is a sum of few-body observables $O_i = \sum_j O_{ij}$, where every qubit is acted on by a constant number of few-body observables

O_{ij}. After N randomized Pauli measurements on copies of ρ with

$$N = \mathcal{O}\left(\frac{\log(\min(M,n))B_\infty^2}{\epsilon^2}\right), \tag{C9}$$

we can estimate $\mathrm{tr}(O_i\rho)$ to ϵ error for all $i \in [M]$ with high probability.

Proof.—The upper bound of $N = \mathcal{O}(\log(M)\max_{i\in[M]}\|O_i\|_\infty^2/\epsilon^2)$ follows immediately from Theorem 10 and Lemma 11. We can also establish an upper bound of $N = \mathcal{O}(\log(n)\max_{i\in[M]}\|O_i\|_\infty^2/\epsilon^2)$. To see this, consider the task of predicting all k-qubit Pauli observables $P \in \{I,X,Y,Z\}^{\otimes n}$ with $|P| \le k$. There are at most $\mathcal{O}(n^k)$ such Pauli observables. To predict all of the k-qubit Pauli observables to ϵ' error under the unknown state ρ, we can combine Theorem 10 and Lemma 9 to see that we need only

$$N = \mathcal{O}\left(\log(n)\max_{i\in[M]}\|O_i\|_\infty^2/(\epsilon')^2\right) \tag{C10}$$

randomized Pauli measurements. Now, given any observable $O_i = \sum_P \alpha_P P$ that is a sum of few-body observables $O_i = \sum_j O_{ij}$, where every qubit is acted on by a constant number of few-body observables O_{ij}, we can predict $\mathrm{tr}(O_i\rho)$ using the identity

$$\mathrm{tr}(O_i\rho) = \sum_{P:\,|P|\le k} \alpha_P\,\mathrm{tr}(P\rho), \tag{C11}$$

which incurs a prediction error of at most $\sum_P |\alpha_P|\epsilon'$. Using the norm inequality in Corollary 12, we have

$$\|O_i\|_{\mathrm{Pauli},1} = \sum_P |\alpha_P| \le C\|O_i\| \tag{C12}$$

for a constant C. Hence, by setting $\epsilon' = \epsilon/C$, we can predict O_i to ϵ error. Thus we can also establish an upper bound of $N = \mathcal{O}(\log(n)\max_{i\in[M]}\|O_i\|_\infty^2/\epsilon^2)$. The claim follows by considering the corresponding prediction algorithm (use the standard classical shadow when $M < n$, and use the above algorithm when $M \ge n$). ■

3. Optimality of Theorem 11

Here we prove the following lower bound on the sample complexity of shadow tomography for bounded-degree observables, demonstrating that Theorem 11 is optimal. The optimality holds even when we consider a collective measurement procedure on many copies of ρ. This is in stark contrast to other sets of observables, such as the collection of high-weight Pauli observables, where single-copy measurements (e.g., classical shadow tomography) require exponentially more copies than collective measurements.

Theorem 12 (Lower bound for predicting bounded-degree observables).—Consider the following task. There is an unknown n-qubit state ρ and we are given M observables $O_1,\ldots,O_M$ with $B_\infty = \max_i\|O_i\|$. Each observable O_i is a sum of few-body observables $O_i = \sum_j O_{ij}$, where every qubit is acted on by a constant number of few-body observables O_{ij}. We would like to estimate $\mathrm{tr}(O_i\rho)$ to ϵ error for all $i \in [M]$ with high probability by performing arbitrary collective measurements on N copies of ρ. The number of copies needs to be at least

$$N = \Omega\left(\frac{\log(\min(M,n))B_\infty^2}{\epsilon^2}\right) \tag{C13}$$

for any algorithm to succeed in this task.

To show Theorem 12, we show a lower bound for the following *distinguishing task*, from which the lower bound for shadow tomography will follow readily. Given $i \in [n]$, let P_i denote the n-body Pauli operator that acts as Z on the ith qubit and trivially elsewhere, and define the mixed state

$$\rho^i \triangleq \frac{1}{2^n}\left(I + \frac{\epsilon}{B_\infty}P_i\right). \tag{C14}$$

We show a lower bound for distinguishing whether ρ is maximally mixed or of the form ρ^i for some i.

Lemma 12 (Lower bound for a distinguishing task).—Let $0 \le \epsilon \le 1$ and $\delta \ge 2\epsilon$. Let $\mathcal{A}$ be an algorithm that, given access to N copies of a mixed state ρ that is either the maximally mixed state or ρ^i for some $i \in [\min(M,n)]$, correctly determines whether or not ρ is maximally mixed with probability at least $3/4$. Then $N = \Omega(\log(\min(M,n))B_\infty^2/\epsilon^2)$.

Proof of Theorem 12.—Let $\mathcal{A}$ be an algorithm that solves the task in Theorem 12 to error $\epsilon/3$. We can use this to give an algorithm for the task in Lemma 12: applying $\mathcal{A}$ to the $\min(M,n)$ observables

$$O_1 \triangleq B_\infty P_1, \quad \ldots, \quad O_{\min(M,n)} \triangleq B_\infty P_{\min(M,n)}, \tag{C15}$$

we can produce $\epsilon/3$-accurate estimates for $\mathrm{tr}(\rho P_j)$ for all $j \in [\min(M,n)]$. Note that if ρ is maximally mixed, $\mathrm{tr}(\rho O_j) = 0$ for all j, whereas if $\rho = \rho^i$ then $\mathrm{tr}(\rho O_j) = \epsilon\mathbb{1}[i=j]$. In particular, by checking whether there is a j for which $\mathrm{tr}(\rho P_j) > 2\epsilon/3$, we can determine whether ρ is maximally mixed or equal to some ρ^i. The lower bound in Lemma 12 thus implies the lower bound in Theorem 12. ■

For convenience, define $n' \triangleq \min(M,n)$. Note that, for any $i \in [n]$, $(\rho^i)^{\otimes N}$ is diagonal, so we can assume without loss of generality that $\mathcal{A}$ simply makes N independent measurements in the computational basis. Proving Lemma 12 thus amounts to showing a lower bound for a classical distribution testing task.

Note that distribution π^i over outcomes of a single measurement of ρ^i in the computational basis places

$$\frac{1+(-1)^{x_i}\epsilon}{2^n} \tag{C16}$$

mass on each string $x \in \{0,1\}^n$. Distribution π over outcomes of a single measurement of the maximally mixed state in the computational basis is uniform over all strings $x \in \{0,1\}^n$. The following basic result in binary hypothesis testing lets us reduce proving Lemma 12 to upper bounding

$$d_{\mathrm{TV}}\Big(\mathop{\mathbb{E}}_i[(\pi^i)^{\otimes N}], \pi^{\otimes N}\Big). \tag{C17}$$

Lemma 13 (Le Cam's two-point method [81]).—Let p_0, p_1 be distributions over a domain Ω for which there exists a distribution D such that $d_{\mathrm{TV}}(p_0, p_1) < 1/3$. Then there is no algorithm $\mathcal{A}$ that maps elements of Ω to $\{0,1\}$ for which $\Pr_{x\sim p_i}[\mathcal{A}(x) = i] \geq 2/3$ for both $i = 0, 1$.

Proof of Lemma 12.—To bound the expression in Eq. (C17), it suffices to bound the chi-squared divergence $\chi^2(\mathbb{E}_i[(\pi^i)^{\otimes N}] \| \pi^{\otimes N})$ because, for any distributions p, q, we have $d_{\mathrm{TV}}(p,q) \leq 2\sqrt{\chi^2(p\|q)}$. For convenience, let us define the likelihood ratio perturbation

$$\eta^i(x) \triangleq \frac{\mathrm{d}\pi^i}{\mathrm{d}\pi}(x) - 1 = (-1)^{x_i}\epsilon, \tag{C18}$$

and observe that, for any $i, j \in [n]$,

$$\mathop{\mathbb{E}}_{x\sim\pi}[\eta^i(x)\eta^j(x)] = \epsilon^2 \mathbb{1}[i=j]. \tag{C19}$$

Also, given strings $x^1, \ldots, x^N \in \{0,1\}^n$ and $S \subseteq [N]$, define

$$\eta^i(x^S) \triangleq \prod_{j\in S} \eta^i(x_j). \tag{C20}$$

We then have the standard calculation (see, e.g., [82, Lemma 22.1])

$$\begin{aligned}
1 + \chi^2\Big(\mathop{\mathbb{E}}_{i\sim[n']}[(\pi^i)^{\otimes N}] \Big\| \pi^{\otimes N}\Big)
&= \mathop{\mathbb{E}}_{x^1,\ldots,x^N\sim\pi^{\otimes N}}\left[\mathop{\mathbb{E}}_{i\sim[n']}\left[\prod_{j=1}^N (1+\eta^{i,t}(x^j))\right]^2\right] \\
&= \mathop{\mathbb{E}}_{i,i'\sim[n']}\left[\mathop{\mathbb{E}}_{x^1,\ldots,x^N\sim\pi^{\otimes N}}\left[\sum_{S,T\subseteq[N]} \eta^i(x^S)\eta^{i'}(x^T)\right]\right] \\
&= \mathop{\mathbb{E}}_{i,i'\sim[n']}\left[\mathop{\mathbb{E}}_{x^1,\ldots,x^N\sim\pi^{\otimes N}}\left[\sum_{S\subseteq[N]} \eta^i(x^S)\eta^{i'}(x^S)\right]\right] \\
&= \mathop{\mathbb{E}}_{i,i'\sim[n']}\left[\mathop{\mathbb{E}}_{x^1,\ldots,x^N\sim\pi^{\otimes N}}\left[\prod_{j=1}^N (1+\eta^i(x^j)\eta^{i'}(x^j))\right]\right] \\
&= \mathop{\mathbb{E}}_{i,i'\sim[n']}\left[\Big(1 + \mathop{\mathbb{E}}_{x\sim\pi}[\eta^i(x)\eta^{i'}(x)]\Big)^N\right] \\
&= \frac{1}{n'}(1+\epsilon^2)^N + \frac{n'-1}{n'}.
\end{aligned} \tag{C21}$$

We conclude that

$$\chi^2\Big(\mathop{\mathbb{E}}_{i\sim[n']}[(\pi^i)^{\otimes N}] \Big\| \pi^{\otimes N}\Big) \leq \frac{1}{n'}((1+\epsilon^2)^N - 1), \tag{C22}$$

so, for $N = c\log(n')/\epsilon^2$ for a sufficiently small constant $c > 0$, this quantity is less than 1/3. By applying Lemma 13 to $p_0 = \pi^{\otimes N}$ and $p_1 = \mathbb{E}_{i\sim[n']}[(\pi^i)^{\otimes N}]$, we obtain the claimed lower bound. ■

APPENDIX D: LEARNING TO PREDICT AN UNKNOWN OBSERVABLE

We begin with a definition of invariance for distribution over quantum states.

Definition 5 (Invariance under a unitary).—A probability distribution $\mathcal{D}$ over quantum states is invariant under a unitary U if the probability density remains unchanged after the action of U, i.e.,

$$f_{\mathcal{D}}(\rho) = f_{\mathcal{D}}(U\rho U^\dagger) \tag{D1}$$

for any state ρ.

In this appendix, we utilize the norm inequalities in Appendix B to give a learning algorithm that achieves the following guarantee. The learning algorithm can learn any unknown n-qubit observable $O^{(\mathrm{unk})}$ even if the scale $\|O^{(\mathrm{unk})}\|$ is unknown. The mean squared error $\mathbb{E}_{\rho\sim\mathcal{D}}|h(\rho) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2$ scales quadratically with the scale of the unknown observable $O^{(\mathrm{unk})}$. We can see that the sample complexity N has a quasipolynomial dependence on the errors ϵ, ϵ' relative to the scale of the unknown observable $O^{(\mathrm{unk})}$, and depends only on the system size n and the failure probability δ logarithmically.

Theorem 13 (Learning to predict an unknown observable).—Let $n, \epsilon, \epsilon', \delta > 0$. Consider any unknown n-qubit observable $O^{(\mathrm{unk})} = \sum_P \alpha_P P$ and any unknown n-qubit state distribution $\mathcal{D}$ that is invariant under single-qubit H and S gates. Suppose that the training data $\{\rho_\ell, \mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell)\}_{\ell=1}^N$ are of size

$$\begin{aligned}
N = \log\left(\frac{n}{\delta}\right)\min(&2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}, \\
&2^{\mathcal{O}[\log(1/\epsilon)\log(n)]}).
\end{aligned} \tag{D2}$$

Let $k = \lceil \log_{1.5}(1/\epsilon) \rceil$, $O^{(\mathrm{low})} = \sum_{|P|\leq k} \alpha_P P$ be the low-degree approximation of $O^{(\mathrm{unk})}$, and $r = 2k/(k+1) \in [1, 2)$. The algorithm can learn a function $h(\rho) = \max(-\hat{\Theta}, \min(\hat{\Theta}, \mathrm{tr}(\hat{O}\rho)))$ for an observable $\hat{O}$ and a real number $\hat{\Theta}$ that achieves a prediction error

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \left(\epsilon + \epsilon'\left[1 + \left(\frac{\|O^{(\mathrm{low})}\|}{\|O^{(\mathrm{unk})}\|}\right)^r\right]\right)\|O^{(\mathrm{unk})}\|^2 \quad \text{(D3)}$$

with probability at least $1 - \delta$.

1. Low-degree approximation under the mean squared error

In order to characterize the mean squared error $\mathbb{E}_{\rho\sim\mathcal{D}}\, \mathrm{tr}(O_1\rho) - \mathrm{tr}(O_2)\rho$ between two observables O_1, O_2, we need the following definition of a modified purity for quantum states.

Definition 6 (Nonidentity purity).—Given a k-qubit state ρ, the nonidentity purity of ρ is

$$\gamma^\star(\rho) \triangleq \frac{1}{2^k} \sum_{Q\in\{X,Y,Z\}^{\otimes k}} \mathrm{tr}(Q\rho)^2. \quad \text{(D4)}$$

Nonidentity purity is bounded by purity:

$$\gamma^\star(\rho) \leq \gamma(\rho) = \mathrm{tr}(\rho^2) = \frac{1}{2^k} \sum_{Q\in\{I,X,Y,Z\}^{\otimes k}} \mathrm{tr}(Q\rho)^2.$$

Lemma 14 (Mean squared error).—Given two n-qubit observables O_1, O_2 with

$$O_1 - O_2 = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \Delta\alpha_P P, \quad \text{(D5)}$$

and a distribution $\mathcal{D}$ over quantum states that is invariant under single-qubit H and S gates, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\mathrm{tr}(O_1\rho) - \mathrm{tr}(O_2\rho)|^2 = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} [\gamma^\star(\rho_{\mathrm{dom}(P)})] \left(\frac{2}{3}\right)^{|P|} |\Delta\alpha_P|^2. \quad \text{(D6)}$$

Proof.—Consider $U_1, \ldots, U_n$ to be independent random single-qubit Clifford unitaries. Because $\mathcal{D}$ is invariant under single-qubit Hadamard and phase gates, $\mathcal{D}$ is invariant under any tensor product of single-qubit Clifford unitaries. This implies that the distribution of the random state ρ is the same as the distribution of the random state $(U_1 \otimes \cdots \otimes U_n)\rho(U_1 \otimes \cdots \otimes U_n)^\dagger$. Using this fact, we expand the mean squared error as

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\mathrm{tr}(O_1\rho) - \mathrm{tr}(O_2\rho)|^2 = \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \mathop{\mathbb{E}}_{U_1,\ldots,U_n} \sum_{P,Q\in\{I,X,Y,Z\}^{\otimes n}} \Delta\alpha_P \Delta\alpha_Q\, \mathrm{tr}\left(\left(\bigotimes_{i=1}^n U_i^\dagger P_i U_i\right) \otimes \left(\bigotimes_{i=1}^n U_i^\dagger Q_i U_i\right)(\rho\otimes\rho)\right). \quad \text{(D7)}$$

Using the unitary 2-design property of a random Clifford unitary and $\mathrm{SWAP} = \frac{1}{2}\sum_{P\in\{I,X,Y,Z\}} P\otimes P$, we have

$$\mathop{\mathbb{E}}_{U_i}\left[U_i^\dagger P_i U_i \otimes U_i^\dagger Q_i U_i\right] = \begin{cases} I \otimes I, & P_i = Q_i = I, \\ \frac{1}{3}(X\otimes X + Y\otimes Y + Z\otimes Z), & P_i = Q_i \neq I, \\ 0, & P_i \neq Q_i. \end{cases} \quad \text{(D8)}$$

We can now write the target value as

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\mathrm{tr}(O_1\rho) - \mathrm{tr}(O_2\rho)|^2 = \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \frac{1}{3^{|P|}} |\Delta\alpha_P|^2 \sum_{Q\in\{X,Y,Z\}^{\otimes|P|}} \mathrm{tr}(Q\rho_{\mathrm{dom}(P)})^2. \quad \text{(D9)}$$

The claim follows from Definition 6 on nonidentity purity $\gamma^\star$. ■

The following lemma tells us that the mean absolute error can be upper bounded by the root-mean-squared error. Hence, both the mean absolute error and the mean squared error are characterized by the ℓ_2 distance between the Pauli coefficients (as well as the average nonidentity purity). Because of the following relation, we focus on the mean squared error throughout the text.

Lemma 15 (Mean absolute error).—Given two n-qubit observables O_1, O_2 with

$$O_1 - O_2 = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \Delta\alpha_P P, \tag{D10}$$

and a distribution $\mathcal{D}$ over quantum states that is invariant under single-qubit H and S gates, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(O_1\rho) - \operatorname{tr}(O_2\rho)| \leq \left(\sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} [\gamma^\star(\rho_{\mathsf{dom}(P)})] \left(\frac{2}{3}\right)^{|P|} |\Delta\alpha_P|^2 \right)^{1/2}. \tag{D11}$$

Proof.—Jensen's inequality gives

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(O_1\rho) - \operatorname{tr}(O_2\rho)| \leq \left(\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(O_1\rho) - \operatorname{tr}(O_2\rho)|^2 \right)^{1/2}. \tag{D12}$$

Combining with Lemma 14 yields the stated result. ■

From Lemma 14, we can construct a low-degree approximation by removing all high-weight Pauli terms for any observable O. The approximation error decays exponentially with the weight of the Pauli terms.

Corollary 13 (Low-degree approximation).—Suppose that we have an n-qubit observable $O = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \alpha_P P$ and a distribution $\mathcal{D}$ over quantum states that is invariant under single-qubit H and S gates. For $k > 0$, consider $O^{(k)} = \sum_{P:\,|P|<k} \alpha_P P$. We have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(O\rho) - \operatorname{tr}(O^{(k)}\rho)|^2 \leq \left(\frac{2}{3}\right)^k \|O\|^2. \tag{D13}$$

Proof.—Using Lemma 14 and the fact that $\gamma^\star(\varrho) \leq \gamma(\varrho) \leq 1$ for any state ϱ, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(O\rho) - \operatorname{tr}(O^{(k)}\rho)|^2 &\leq \sum_{P:\,|P|\geq k} \left(\frac{2}{3}\right)^{|P|} |\alpha_P|^2 \\ &\leq \left(\frac{2}{3}\right)^k \sum_P |\alpha_P|^2. \end{aligned} \tag{D14}$$

The norm inequality given in Proposition 1 establishes the claim. ■

2. Tools for extracting and filtering Pauli coefficients

In order to learn the low-degree approximation of an arbitrary observable O, we need to be able to extract the relevant α_P. Furthermore, we impose criteria for filtering out uninfluential Pauli observables P to prevent them from increasing the noise and leading to a higher prediction error.

a. Extracting the Pauli coefficient

Lemma 16 (Extracting the Pauli coefficient).—Suppose that we have an n-qubit observable $O = \sum_{P\in\{I,X,Y,Z\}^{\otimes n}} \alpha_P P$ and a distribution $\mathcal{D}$ over quantum states that is invariant under single-qubit H and S gates. For any Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \operatorname{tr}(O\rho)\operatorname{tr}(P\rho) = \left(\frac{2}{3}\right)^{|P|} \alpha_P \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \gamma^*(\rho_{\mathsf{dom}(P)}). \tag{D15}$$

Proof.—Using the invariance of $\mathcal{D}$, we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \operatorname{tr}(O\rho)\operatorname{tr}(P\rho) \\ &\quad= \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \mathop{\mathbb{E}}_{U_1,\ldots,U_n} \sum_{Q\in\{I,X,Y,Z\}^{\otimes n}} \alpha_Q \operatorname{tr}\left(\left(\bigotimes_{i=1}^n U_i^\dagger P_i U_i \right) \right. \\ &\quad\quad \left. \otimes \left(\bigotimes_{i=1}^n U_i^\dagger Q_i U_i \right) (\rho\otimes\rho) \right). \end{aligned} \tag{D16}$$

Using Eq. (D8), we can rewrite the above expression as

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \operatorname{tr}(O\rho)\operatorname{tr}(P\rho) = \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \frac{1}{3^{|P|}} \alpha_P \sum_{Q\in\{X,Y,Z\}^{\otimes |P|}} \operatorname{tr}(Q\rho_{\mathsf{dom}(P)})^2. \tag{D17}$$

The claim follows from the definition of the nonidentity purity γ^*. ■

For each Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$, define the quantity we can extract using the lemma to be

$$x_P = \left(\frac{2}{3}\right)^{|P|} \alpha_P \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \gamma^*(\rho_{\mathsf{dom}(P)}). \tag{D18}$$

We can obtain an estimate $\hat{x}_P$ for x_P by averaging $\operatorname{tr}(O\rho)\operatorname{tr}(P\rho)$ over the training data. However, to obtain an estimate $\hat{\alpha}_P$ for α_P, we need to divide $\hat{x}$ by $(2/3)^{|P|}\mathbb{E}_{\rho\sim\mathcal{D}}\gamma^*(\rho_{\mathsf{dom}(P)})$. The error in the estimate $\hat{\alpha}_P$ could be arbitrarily large if $(2/3)^{|P|}\mathbb{E}_{\rho\sim\mathcal{D}}\gamma^*(\rho_{\mathsf{dom}(P)})$ is close to zero. Hence, we present a filter in Appendix D 2 b below to handle this issue. In addition to this filter, the norm inequalities given in Appendix B show that most α_P would be close to zero. Hence, when α_P is small, we could simply set them to zero to avoid noise build-up. This gives rise to the second filtering layer given in Appendix D 2 c below.

b. Filtering the small-weight factor

The first filter sets the estimate $\hat{\alpha}_P$ to be zero when the average nonidentity purity $\mathbb{E}_{\rho\sim\mathcal{D}}\gamma^*(\rho_{\mathsf{dom}(P)})$ is close to zero. We define the weight factor for a Pauli observable P to be

$$\beta_P = \left(\frac{2}{3}\right)^{|P|} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} \gamma^*(\rho_{\mathsf{dom}(P)}). \tag{D19}$$

The weight factor β_P depends on distribution $\mathcal{D}$, which may be unknown. Hence, we can only obtain an estimate $\hat{\beta}_P$ for β_P by utilizing the training data. Recall from Lemma 16 that we can only obtain an estimate $\hat{x}_P$ for $x_P = \alpha_P\beta_P$. The mean squared error (Lemma 14) shows that the contribution from error in $\hat{\alpha}_P$ is

$$\beta_P|\hat{\alpha}_P - \alpha_P|^2. \tag{D20}$$

The presence of β_P in the mean squared error is very useful since it counteracts the fact that we cannot estimate $\hat{\alpha}_P$ accurately when β_P is close to zero. The following lemma shows that estimates for β_P and x_P are sufficient to perform filtering and achieve a small mean squared error.

Lemma 17 (Filtering the small-weight factor).—Let $\tilde{\epsilon}, \eta > 0$. Consider $\alpha \in [-\eta, \eta]$ and $\beta \in [0, 1]$. Let $x = \alpha\beta \in [-\eta, \eta]$. Given estimates $\hat{x}$ and $\hat{\beta}$ with $|\hat{x} - x| < \eta\tilde{\epsilon}$ and $|\hat{\beta} - \beta| < \tilde{\epsilon}$, if we define the estimate

$$\hat{\alpha} = \begin{cases} 0, & \hat{\beta} \le 2\tilde{\epsilon}, \\ \hat{x}/\hat{\beta}, & \hat{\beta} > 2\tilde{\epsilon}, \end{cases} \tag{D21}$$

then we have $\beta|\hat{\alpha} - \alpha|^2 \le 3\eta^2\tilde{\epsilon}$.

Proof.—Consider the first case in which $\hat{\beta} \le 2\tilde{\epsilon}$. We have

$$\beta|\hat{\alpha} - \alpha|^2 = \beta\alpha^2 \le \eta^2\beta \le \eta^2\hat{\beta} + \eta^2\tilde{\epsilon} \le 3\eta^2\tilde{\epsilon}. \tag{D22}$$

For the second case in which $\hat{\beta} > 2\tilde{\epsilon}$, we have $\beta > \tilde{\epsilon}$. By applying the triangle inequality, we have

$$|\sqrt{\beta}\hat{\alpha} - \sqrt{\beta}\alpha| \le \frac{\sqrt{\beta}}{\hat{\beta}}|\hat{x} - x| + |\sqrt{\beta}x|\left|\frac{1}{\hat{\beta}} - \frac{1}{\beta}\right|. \tag{D23}$$

The first term can be bounded as $\sqrt{\beta}|\hat{x} - x|/\hat{\beta} \le \eta\sqrt{\beta}\tilde{\epsilon}/\hat{\beta}$. The second term can be bounded by the same expression

$$|\sqrt{\beta}x|\left|\frac{1}{\hat{\beta}} - \frac{1}{\beta}\right| = \beta^{3/2}|\alpha|\frac{|\hat{\beta} - \beta|}{\hat{\beta}\beta} \le \eta\frac{\sqrt{\beta}}{\hat{\beta}}\tilde{\epsilon}. \tag{D24}$$

Using the fact that $\sqrt{z + \tilde{\epsilon}}/z$ is monotonically decreasing for $z > 0$, we have

$$\frac{\sqrt{\beta}}{\hat{\beta}}\tilde{\epsilon} \le \frac{\sqrt{\hat{\beta} + \tilde{\epsilon}}}{\hat{\beta}}\tilde{\epsilon} \le \sqrt{\frac{3}{4}}\tilde{\epsilon}. \tag{D25}$$

Together, $|\sqrt{\beta}\hat{\alpha} - \sqrt{\beta}\alpha|^2 \le 3\eta^2\tilde{\epsilon}$ and the claim is established. ∎

c. Filtering uninfluential Pauli observables

Consider a set $S \subseteq \{I, X, Y, Z\}^{\otimes n}$ that contains the Pauli observables of interest. For example, we later consider S to be the set of all few-body Pauli observables. Using the norm inequalities given in Appendix B, we can filter out more α_P to achieve an improved mean squared error. Below is the filtering lemma that combines both the filtering of Pauli observables with a small weight factor (Lemma 17) and the filtering of those with a small contribution (characterized by $|x_P|/\beta_P^{1/2}$).

Lemma 18 (Filtering lemma).—Suppose that $\tilde{\epsilon}, \eta > 0$ and that we have a set $S \subseteq \{I, X, Y, Z\}^{\otimes n}$. Consider $\alpha_P \in [-\eta, \eta]$, $\beta_P \in [0, 1]$, and $x_P = \alpha_P\beta_P \in [-\eta, \eta]$ for all $P \in S$. Assume that there exist $A > 0$ and $1 \le r < 2$ such that

$$\sum_{P\in S} |\alpha_P|^r \le A^r. \tag{D26}$$

Given $\hat{x}_P$ and $\hat{\beta}_P$ with $|\hat{x}_P - x_P| < \eta\tilde{\epsilon}$ and $|\hat{\beta}_P - \beta_P| < \tilde{\epsilon}$ for all $P \in S$, if we define

$$\hat{\alpha}_P = \begin{cases} 0, & \hat{\beta}_P \le 2\tilde{\epsilon}, \\ 0, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P|/\hat{\beta}_P^{1/2} \le 2\eta\sqrt{\tilde{\epsilon}}, \\ \hat{x}_P/\hat{\beta}_P, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P|/\hat{\beta}_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}}, \end{cases} \tag{D27}$$

then we have $\sum_{P\in S} \beta_P|\hat{\alpha}_P - \alpha_P|^2 \le 6A^r\eta^{2-r}\tilde{\epsilon}^{1-(r/2)}$. We also have $\beta_P|\hat{\alpha}_P - \alpha_P|^2 \le 9\eta^2\tilde{\epsilon}$ for all $P \in S$.

Proof.—We first define $S^u \subseteq S$ to be the set of Pauli observables P with $\hat{\beta}_P > 2\tilde{\epsilon}$, $|\hat{x}_P|/\hat{\beta}_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}}$. The set S^u contains all the unfiltered Pauli observables. We define S^f to be $S \setminus S^u$, which contains all the filtered Pauli observables. We separate the contributions of S^u and S^f in the mean squared error $\sum_{P \in S} \beta_P |\hat{\alpha}_P - \alpha_P|^2$:

$$\sum_{P \in S} \beta_P |\hat{\alpha}_P - \alpha_P|^2 = \sum_{P \in S^u} \beta_P |\hat{\alpha}_P - \alpha_P|^2 + \sum_{P \in S^f} \beta_P |\hat{\alpha}_P - \alpha_P|^2. \tag{D28}$$

A key quantity for the analysis is $\beta_P^{1/2}\alpha_P = x_P/\beta_P^{1/2}$. For Pauli P with $\hat{\beta}_P \leq 2\tilde{\epsilon}$, we have

$$|\beta_P^{1/2}\alpha_P| \leq \eta\sqrt{\hat{\beta}_P + \tilde{\epsilon}} \leq \eta\sqrt{3\tilde{\epsilon}}. \tag{D29}$$

For Pauli P with $\hat{\beta}_P > 2\tilde{\epsilon}$, we have

$$\begin{aligned} \left|\frac{\hat{x}_P}{\hat{\beta}_P^{1/2}} - \frac{x_P}{\beta_P^{1/2}}\right| &\leq \frac{1}{\hat{\beta}_P^{1/2}}|\hat{x}_P - x_P| + |x_P|\left|\frac{1}{\hat{\beta}_P^{1/2}} - \frac{1}{\beta_P^{1/2}}\right| \\ &\leq \eta\sqrt{\frac{\tilde{\epsilon}}{2}} + \eta\left|\frac{\beta_P}{\hat{\beta}_P^{1/2}} - \beta_P^{1/2}\right| \\ &\leq \eta\sqrt{\tilde{\epsilon}}. \end{aligned} \tag{D30}$$

The last inequality uses the fact that $\beta_P > \tilde{\epsilon}$, $\hat{\beta}_P/\beta_P > 2$, and, hence,

$$\left|\frac{\beta_P}{\hat{\beta}_P^{1/2}} - \beta_P^{1/2}\right| = \frac{|\hat{\beta}_P - \beta_P|}{\hat{\beta}_P^{1/2}(1 + (\hat{\beta}_P/\beta_P)^{1/2})} \leq \frac{\sqrt{\tilde{\epsilon}}}{2 + \sqrt{2}}. \tag{D31}$$

We are now ready to analyze the contributions of S^u and S^f.

For the unfiltered Pauli observables (those in set S^u), we can use Lemma 17 to obtain

$$\sum_{P \in S^u} \beta_P |\hat{\alpha}_P - \alpha_P|^2 \leq 3\eta^2\tilde{\epsilon}|S^u|. \tag{D32}$$

Equation (D30) shows that, for Pauli observable P with $\hat{\beta}_P > 2\tilde{\epsilon}$ and $|\hat{x}_P|/\hat{\beta}_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}}$, we have $|x_P|/\beta_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}} - \eta\sqrt{\tilde{\epsilon}}$. We use this fact to bound the size of set $|S^u|$:

$$\begin{aligned} |S^u| &\leq \sum_{P \in S^u} \frac{(|x_P|/\beta_P^{1/2})^r}{(2\eta\sqrt{\tilde{\epsilon}} - \eta\sqrt{\tilde{\epsilon}})^r} = \frac{1}{\eta^r\tilde{\epsilon}^{r/2}} \sum_{P \in S^u} |\alpha_P|^r \beta_P^{r/2} \\ &\leq \frac{1}{\eta^r\tilde{\epsilon}^{r/2}} \sum_{P \in S} |\alpha_P|^r = \frac{A^r}{\eta^r\tilde{\epsilon}^{r/2}}. \end{aligned} \tag{D33}$$

Together, we have the upper bound

$$\sum_{P \in S^u} \beta_P |\hat{\alpha}_P - \alpha_P|^2 \leq 3\eta^{2-r}A^r\tilde{\epsilon}^{1-r/2}. \tag{D34}$$

For the filtered Pauli observables (those in set S^f), we have

$$\sum_{P \in S^f} \beta_P |\hat{\alpha}_P - \alpha_P|^2 = \sum_{P \in S^f} |\beta_P^{1/2}\alpha_P|^r |\beta_P^{1/2}\alpha_P|^{2-r}. \tag{D35}$$

There are two types of Pauli observables in S^f.

(1) For P with $\hat{\beta}_P \leq 2\tilde{\epsilon}$, we have $|\beta_P^{1/2}\alpha_P| \leq \eta\sqrt{3\tilde{\epsilon}}$ from Eq. (D29).
(2) For P with $\hat{\beta}_P > 2\tilde{\epsilon}$ and $|\hat{x}_P|/\hat{\beta}_P^{1/2} \leq \eta 2\sqrt{\tilde{\epsilon}}$, we have $|\beta_P^{1/2}\alpha_P| = |x_P|/\beta_P^{1/2} \leq 2\eta\sqrt{\tilde{\epsilon}} + \eta\sqrt{\tilde{\epsilon}}$ from Eq. (D30).

Together, we have the upper bound

$$\begin{aligned} \sum_{P \in S^f} \beta_P |\hat{\alpha}_P - \alpha_P|^2 &\leq (3\eta\sqrt{\tilde{\epsilon}})^{2-r} \sum_{P \in S^f} \beta_P^{r/2}|\alpha_P|^r \\ &\leq A^r(3\eta\sqrt{\tilde{\epsilon}})^{2-r} \\ &\leq 3A^r\eta^{2-r}\tilde{\epsilon}^{1-r/2}. \end{aligned} \tag{D36}$$

Combining the contributions of S^u and S^f yields

$$\sum_{P \in S} \beta_P |\hat{\alpha}_P - \alpha_P|^2 \leq 6A^r\eta^{2-r}\tilde{\epsilon}^{1-r/2}. \tag{D37}$$

Thus we have established the first statement of the lemma.

We now focus on the second statement of the lemma. For Pauli observable P that satisfies the first and the third cases of Eq. (D27), we can use Lemma 17 to obtain $\beta_P|\hat{\alpha}_P - \alpha_P|^2 \leq 3\eta^2\tilde{\epsilon} < 9\eta^2\tilde{\epsilon}$. For the second case of Eq. (D27), we can use Eq. (D30) to see that

$$\begin{aligned} \beta_P|\hat{\alpha}_P - \alpha_P|^2 &= \left(\frac{x_P}{\beta_P^{1/2}}\right)^2 \leq \left(\frac{|\hat{x}_P|}{\hat{\beta}_P^{1/2}} + \left|\frac{\hat{x}_P}{\hat{\beta}_P^{1/2}} - \frac{x_P}{\beta_P^{1/2}}\right|\right)^2 \\ &\leq 9\eta^2\tilde{\epsilon}. \end{aligned} \tag{D38}$$

Hence, for all $P \in S$, we have $\beta_P|\hat{\alpha}_P - \alpha_P|^2 \leq 9\eta^2\tilde{\epsilon}$. ■

3. Learning algorithm

In this section, we present a learning algorithm satisfying the guarantee given in Theorem 13. Consider the full training data $\{\rho_\ell, y_\ell = \mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell)\}_{\ell=1}^N$ of size N. The learning algorithm splits the full data into a smaller training set of size N_{tr} and a validation set of size N_{val} with $N = N_{\mathrm{tr}} + N_{\mathrm{val}}$. The training set is used to extract Pauli coefficients and perform filtering with a hyperparameter η. The validation set is used to choose the best hyperparameter η. We can set $N_{\mathrm{tr}} = (4/5)N$ and $N_{\mathrm{val}} = (1/5)N$.

We consider two slightly different learning algorithms for the sample complexity scalings of

$$N = \log\left(\frac{n}{\delta}\right) 2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}} \quad \text{and}$$

$$N = \log\left(\frac{n}{\delta}\right) 2^{\mathcal{O}[\log(1/\epsilon)\log(n)]}. \tag{D39}$$

We can simply look at which sample complexity is smaller and select the corresponding learning algorithm.

We begin with the learning algorithm for achieving the sample complexity on the left of Eq. (D39). First, the algorithm computes the sample maximum over the training set,

$$\hat{\Theta} = \max_{\ell \in \{1,\ldots,N_{\mathrm{tr}}\}} |y_\ell| = \max_{\ell \in \{1,\ldots,N_{\mathrm{tr}}\}} |\,\mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell))| \leq \|O^{(\mathrm{unk})}\|, \tag{D40}$$

to obtain a scale for the function value. Let $C(k)$ be the constant from Corollary 11. We define

$$\tilde{\epsilon} \triangleq \left(\frac{\epsilon'}{12}\right)^{k+1} \left(\frac{C(k)}{3}\right)^{2k}. \tag{D41}$$

Next, we consider the grid of hyperparameters

$$\eta \in \{2^0\hat{\Theta}, 2^1\hat{\Theta}, 2^2\hat{\Theta}, \ldots, 2^R\hat{\Theta}\}, \tag{D42}$$

where $R = \log_2\lceil 1/\tilde{\epsilon}\rceil$. For each hyperparameter η, the learning algorithm runs as follows. The learning algorithm considers every Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$ with $|P| \leq \log_{1.5}(1/\epsilon)$. We define the set that contains the Pauli observables of interest,

$$S = \{P\colon |P| \leq \log_{1.5}(1/\epsilon)\}, \tag{D43}$$

and $k = \lceil \log_{1.5}(1/\epsilon)\rceil$. For each $P \in S$, the algorithm computes

$$\hat{x}_P = \frac{1}{N_{\mathrm{tr}}} \sum_{\ell=1}^{N_{\mathrm{tr}}} \mathrm{tr}(P\rho_\ell) y_\ell, \tag{D44}$$

$$\hat{\beta}_P = \frac{1}{N_{\mathrm{tr}}} \sum_{\ell=1}^{N_{\mathrm{tr}}} \mathrm{tr}(P\rho_\ell)\,\mathrm{tr}(P\rho_\ell), \tag{D45}$$

using the training set $\{(\rho_\ell, y_\ell = \mathrm{tr}(O^{\mathrm{unk}\rho_\ell}))\}_{\ell=1}^{N_{\mathrm{tr}}}$. By the definitions of $\hat{x}_P$ and $\hat{\Theta}$, we have

$$|\hat{x}_P| \leq \hat{\Theta} \quad \text{for all } P \in S. \tag{D46}$$

Then, for each $P \in S$, the algorithm computes

$$\hat{\alpha}_P(\eta) = \begin{cases} 0, & \hat{\beta}_P \leq 2\tilde{\epsilon}, \\ 0, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P|/\hat{\beta}_P^{1/2} \leq 2\eta\sqrt{\tilde{\epsilon}}, \\ \hat{x}_P/\hat{\beta}_P, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P|/\hat{\beta}_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}}. \end{cases} \tag{D47}$$

The algorithm considers the function $h(\rho; \eta) = \max(-\hat{\Theta}, \min(\hat{\Theta}, \mathrm{tr}(\hat{O}(\eta)\rho)))$, where the observable $\hat{O}(\eta)$ is defined as

$$\hat{O}(\eta) = \sum_{P \in S} \hat{\alpha}_P(\eta) P. \tag{D48}$$

The best η is selected using the validation set:

$$\eta^* = \operatorname*{arg\,min}_{\eta \in \{2^0\hat{\Theta}, \ldots, 2^R\hat{\Theta}\}} \frac{1}{N_{\mathrm{val}}} \sum_{\ell=N_{\mathrm{tr}}+1}^{N_{\mathrm{tr}}+N_{\mathrm{val}}} |h(\rho_\ell; \eta) - y_\ell|^2. \tag{D49}$$

The learning algorithm outputs $h(\rho; \eta^*)$ as the learned function.

We now present the learning algorithm for achieving the sample complexity on the right of Eq. (D39). We define the set that contains the Pauli observables of interest,

$$S' = \{P\colon |P| \leq \log_{1.5}(2/\epsilon)\}, \tag{D50}$$

and $k' = \lceil \log_{1.5}(2/\epsilon)\rceil$. For each $P \in S'$, the algorithm computes

$$\hat{x}'_P = \frac{1}{N} \sum_{\ell=1}^{N} \mathrm{tr}(P\rho_\ell) y_\ell, \tag{D51}$$

$$\hat{\beta}'_P = \frac{1}{N} \sum_{\ell=1}^{N} \mathrm{tr}(P\rho_\ell)\,\mathrm{tr}(P\rho_\ell), \tag{D52}$$

using the full dataset $\{(\rho_\ell, y_\ell = \mathrm{tr}(O^{\mathrm{unk}\rho_\ell}))\}_{\ell=1}^{N}$. The algorithm uses the hyperparameter

$$\tilde{\epsilon}' \triangleq \frac{\epsilon}{6n^{k'}}. \tag{D53}$$

Then, for each $P \in S'$, the algorithm computes

$$\hat{\alpha}'_P = \begin{cases} 0, & \hat{\beta}'_P \leq 2\tilde{\epsilon}', \\ \hat{x}'_P/\hat{\beta}'_P, & \hat{\beta}'_P > 2\tilde{\epsilon}'. \end{cases} \tag{D54}$$

The algorithm outputs the function $h'(\rho) = \mathrm{tr}(\hat{O}'\rho)$, where the observable $\hat{O}'$ is defined as $\hat{O}' = \sum_{P \in S'} \hat{\alpha}'_P P$.

Here, we assume that $\mathrm{tr}(P\rho_\ell)$ can be obtained from the training data. However, for each $\mathrm{tr}(P\rho_\ell)$, we only need to be able to obtain an unbiased estimator for $\mathrm{tr}(P\rho_\ell)$ and for $\mathrm{tr}(P\rho_\ell)^2$. Recall that an unbiased estimator for *a* is a random variable with expectation value equal to a. For example, an unbiased estimator for $\mathrm{tr}(P\rho_\ell)^2$ can be obtained by performing two quantum measurements on two individual copies of ρ_ℓ using the observable P and multiplying the results, or by utilizing classical shadow formalism [46] and randomized measurement [47].

4. Rigorous performance guarantee

In this section, we prove that the learning algorithm presented in the last section satisfies Theorem 13. We separate the proof for achieving the sample complexity on the left and right of Eq. (D39).

The proof for the sample complexity stated on the left of Eq. (D39) consists of three parts: (1) a characterization of the prediction error, (2) the existence of a good hyperparameter $\eta^\triangle$ that achieves a small prediction error, (3) the best hyperparameter η^* found by a grid search over the validation set must yield a small prediction error.

The proof for the sample complexity stated on the right of Eq. (D39) is simpler and is given at the end.

a. Characterization of the prediction error

We begin with a lemma about the sample maximum.

Lemma 19 (Sample maximum).—Let $1 > \epsilon, \delta > 0$. Consider an arbitrary real-valued random variable X. Let $X_1, \ldots, X_N$ be N independent samples of X with $N = \lceil \log(1/\delta)/\epsilon \rceil$, and let $\hat{\Theta} = \max_i X_i$. Then

$$\Pr[X \leq \hat{\Theta}] \geq 1 - \epsilon \tag{D55}$$

with probability at least $1 - \delta$.

Proof.—Recall that the cumulative distribution function is defined as $F(\theta) = \Pr[X \leq \theta]$. We define the approximate maximum as

$$\Theta \triangleq \inf_{\theta\,:\,F(\theta) \geq 1-\epsilon} \theta. \tag{D56}$$

Using the right continuity of $F(\theta) = \Pr[X \leq \theta]$, we have

$$F(\Theta) = \Pr[X \leq \Theta] \geq 1 - \epsilon. \tag{D57}$$

Furthermore, from the definition of Θ, we have

$$\Pr[X \geq \Theta] \geq \epsilon. \tag{D58}$$

To see the above inequality, suppose that $\Pr[X \geq \Theta] < \epsilon$. Then from the left continuity of $F'(\theta) = \Pr[X \geq \theta]$, we can find $\Theta' < \Theta$ such that $\Pr[X \geq \Theta'] \leq \epsilon$. Thus, there exists $\Theta' < \Theta$ with $\Pr[X \leq \Theta'] \geq 1 - \epsilon$, which is a contradiction to the definition of Θ. Together, we have

$$\Pr[X_i < \Theta \text{ for all } i \in [N]] \leq (1-\epsilon)^N. \tag{D59}$$

By choosing $N = \lceil \log(1/\delta)/\epsilon \rceil$, we have

$$\Pr\left[\max_i X_i \geq \Theta\right] \geq 1 - (1-\epsilon)^{\log(1/\delta)/\epsilon} \geq 1 - \delta. \tag{D60}$$

Thus, with probability at least $1 - \delta$, we have $\hat{\Theta} \geq \Theta$. Using the monotonicity of $F(\theta)$, we have

$$\Pr[X \leq \hat{\Theta}] = F(\hat{\Theta}) \geq F(\Theta) \geq 1 - \epsilon, \tag{D61}$$

which establishes this lemma. ■

Using the above lemma, we can show that, given a training set of size

$$N_{\mathrm{tr}} \geq \frac{12\log(3/\delta)}{\epsilon'}, \tag{D62}$$

the real value $\hat{\Theta} \leq \|O^{(\mathrm{unk})}\|$ obtained by the algorithm satisfies

$$\Pr_{\rho\sim\mathcal{D}}[|\,\mathrm{tr}(O^{(\mathrm{unk})}\rho)| \leq \hat{\Theta}] \geq 1 - \frac{\epsilon'}{12} \tag{D63}$$

with probability at least $1 - \delta/3$. Hence, with probability at least $1 - \delta/3$, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho;\eta) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\,\mathrm{tr}(\hat{O}(\eta)\rho) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 + \frac{\epsilon'}{12}|\hat{\Theta} + \|O^{(\mathrm{unk})}\||^2. \tag{D64}$$

Using Lemma 14 on the mean squared error and Corollary 13 on the low-degree approximation, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} &|h(\rho;\eta) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \\ &\leq \underbrace{(2/3)^k\|O^{(\mathrm{unk})}\|^2}_{\leq\|O^{(\mathrm{unk})}\|^2\epsilon} + \sum_{P\in S} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})] \\ &\times \left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(\eta) - \alpha_P|^2 + \frac{\epsilon'}{3}\|O^{(\mathrm{unk})}\|^2\end{aligned} \tag{D65}$$

with probability at least $1 - \delta/3$.

Let us define the variables

$$x_P \triangleq \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}\alpha_P,$$
$$\beta_P \triangleq \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}, \quad \text{for all } P \in S. \tag{D66}$$

Then, with probability at least $1 - \delta/3$ over the sampling of the training set, we have the following characterization of the prediction error for all $\eta > 0$:

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho;\eta) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \le \epsilon\|O^{(\mathrm{unk})}\|^2 + \frac{\epsilon'}{3}\|O^{(\mathrm{unk})}\|^2 + \sum_{P\in S}\beta_P|\hat{\alpha}_P(\eta) - \alpha_P|^2. \tag{D67}$$

We utilize this form to show the existence of a good hyperparameter $\eta^\triangle$.

b. Existence of a good hyperparameter $\eta^\triangle$

By considering the training set size to be

$$N_{\mathrm{tr}} = \Omega\left(\frac{\log(1/\delta)}{\epsilon'} + \frac{\log(|S|/\delta)}{\tilde{\epsilon}^2}\right), \tag{D68}$$

we can guarantee Eq. (D67) with probability at least $1 - \delta/3$. Furthermore, utilizing Hoeffding's inequality and the union bound, we can also guarantee that

$$|\hat{x}_P - x_P| \le \|O^{(\mathrm{unk})}\|\tilde{\epsilon}, \quad |\hat{\beta}_P - \beta_P| \le \tilde{\epsilon}, \quad \text{for all } P \in S \tag{D69}$$

with probability at least $1 - \delta/3$. The norm inequality given in Corollary 11 shows that

$$\sum_{P\in S}|\alpha_P|^r \le \left(\frac{3}{C(k)}\right)^r \|O^{(\mathrm{low})}\|^r \tag{D70}$$

for a constant given by

$$C(k) = \frac{\sqrt{2(k!)}}{2k^{k+1.5+(k+1)/(2k)}(\sqrt{6}+2\sqrt{3})^k}. \tag{D71}$$

We now condition on the event that Eqs. (D67) and (D69) both hold, which happens with probability at least $1 - (2/3)\delta$. We are now ready to define the good hyperparameter $\eta^\triangle$.

Let hyperparameter $\eta^\triangle$ belonging to the grid in Eq. (D42) be defined as

$$\eta^\triangle = 2^{\min(R, \lceil\log_2(\|O^{(\mathrm{unk})}\|/\hat{\Theta})\rceil)}\hat{\Theta}. \tag{D72}$$

We separately consider two cases: (1) $\eta^\triangle = 2^R\hat{\Theta}$, (2) $\eta^\triangle < 2^R\hat{\Theta}$. For the first case $\eta^\triangle = 2^R\hat{\Theta}$, we can use $|\hat{x}_P| \le \hat{\Theta}$ in Eq. (D46) and the definition of R to see that

$$\hat{\alpha}_P(\eta^\triangle) = 0 \quad \text{for all } P \in S. \tag{D73}$$

Since $\eta^\triangle = 2^R\hat{\Theta}$, we have $R \le \lceil\log_2(\|O^{(\mathrm{unk})}\|/\hat{\Theta})\rceil$. This yields $\eta^\triangle \le 2\|O^{(\mathrm{unk})}\|$, which implies that

$$\hat{\alpha}_P(2\|O^{(\mathrm{unk})}\|) = 0 \quad \text{for all } P \in S. \tag{D74}$$

Hence, the reconstructed Pauli coefficients $\hat{\alpha}_P(\cdot)$ are the same for $\eta^\triangle$ and $2\|O^{(\mathrm{unk})}\|$. The filtering lemma given in Lemma 18 shows that

$$\begin{aligned}\sum_{P\in S}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(\eta^\triangle) - \alpha_P|^2 &= \sum_{P\in S}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(2\|O^{(\mathrm{unk})}\|) - \alpha_P|^2 \\ &\le 12\left(\frac{3}{C(k)}\right)^r\|O^{(\mathrm{unk})}\|^{2-r}\|O^{(\mathrm{low})}\|^r\tilde{\epsilon}^{1-r/2}.\end{aligned} \tag{D75}$$

For the second case $\eta^\triangle < 2^R\hat{\Theta}$, we have the following bound on $\eta^\triangle$:

$$\eta^\triangle = 2^{\lceil\log_2(\|O^{(\mathrm{unk})}\|/\hat{\Theta})\rceil}\hat{\Theta} \in [\|O^{(\mathrm{unk})}\|, 2\|O^{(\mathrm{unk})}\|]. \tag{D76}$$

The filtering lemma given in Lemma 18 shows that

$$\begin{aligned}\sum_{P\in S}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(\eta^\triangle) - \alpha_P|^2 &\le 6(\eta^\triangle)^r\left(\frac{3}{C(k)}\right)^r\|O^{(\mathrm{low})}\|^r\tilde{\epsilon}^{1-r/2} \\ &\le 12\left(\frac{3}{C(k)}\right)^r\|O^{(\mathrm{unk})}\|^{2-r}\|O^{(\mathrm{low})}\|^r\tilde{\epsilon}^{1-r/2}.\end{aligned} \tag{D77}$$

In both cases (1) and (2), using the definitions $r = 2k/(k+1)$ and $\tilde{\epsilon} = (\epsilon'/12)^{k+1}(C(k)/3)^{2k}$, we have

$$\sum_{P\in S}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(\eta^\triangle) - \alpha_P|^2 \le \epsilon'\|O^{(\mathrm{unk})}\|^{2-r}\|O^{(\mathrm{low})}\|^r. \tag{D78}$$

Combining with Eq. (D67), we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho;\eta^\triangle) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \le \epsilon\|O^{(\mathrm{unk})}\|^2 + \frac{\epsilon'}{3}\|O^{(\mathrm{unk})}\|^2 + \epsilon'\|O^{(\mathrm{unk})}\|^{2-r}\|O^{(\mathrm{low})}\|^r \tag{D79}$$

with probability at least $1 - (2/3)\delta$.

c. The prediction performance of hyperparameter η^*

From the definition of $h(\rho;\eta)$, for any quantum state ρ, we have

$$|h(\rho;\eta) - \mathrm{tr}(O^{(\mathrm{unk})}\rho))|^2 \le |\hat{\Theta} + \|O^{(\mathrm{unk})}\||^2 \le 4\|O^{(\mathrm{unk})}\|^2. \tag{D80}$$

Using Hoeffding's inequality and the union bound, we can show that, given a validation set of size

$$N_{\mathrm{val}} = \Omega\left(\frac{\log(R/\delta)}{(\epsilon')^2}\right) \tag{D81}$$

with probability at least $1-\delta/3$, we have

$$\left|\frac{1}{N_{\mathrm{val}}}\sum_{\ell=N_{\mathrm{tr}}+1}^{N_{\mathrm{tr}}+N_{\mathrm{val}}}|h(\rho_\ell;\eta)-\mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell))|^2-\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho;\eta)-\mathrm{tr}(O^{(\mathrm{unk})}\rho))|^2\right|\le\|O^{(\mathrm{unk})}\|^2\frac{\epsilon'}{3} \tag{D82}$$

for all $\eta\in\{2^0\hat{\Theta},\dots,2^R\hat{\Theta}\}$. Using the definitions of η^* and $\eta^\triangle$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho;\eta^*)-\mathrm{tr}(O^{(\mathrm{unk})}\rho))|^2&\le\frac{1}{N_{\mathrm{val}}}\sum_{\ell=N_{\mathrm{tr}}+1}^{N_{\mathrm{tr}}+N_{\mathrm{val}}}|h(\rho_\ell;\eta^*)-\mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell))|^2+\|O^{(\mathrm{unk})}\|^2\frac{\epsilon'}{3}\\&\le\frac{1}{N_{\mathrm{val}}}\sum_{\ell=N_{\mathrm{tr}}+1}^{N_{\mathrm{tr}}+N_{\mathrm{val}}}|h(\rho_\ell;\eta^\triangle)-\mathrm{tr}(O^{(\mathrm{unk})}\rho_\ell))|^2+\|O^{(\mathrm{unk})}\|^2\frac{\epsilon'}{3}\\&\le\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho;\eta^\triangle)-\mathrm{tr}(O^{(\mathrm{unk})}\rho))|^2+\|O^{(\mathrm{unk})}\|^2\frac{2\epsilon'}{3}\end{aligned} \tag{D83}$$

with probability at least $1-\delta/3$ over the sampling of the validation set. Combining with Eq. (D79) and employing the union bound, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho;\eta^*)-\mathrm{tr}(O^{(\mathrm{unk})}\rho))|^2\le\epsilon\|O^{(\mathrm{unk})}\|^2+\epsilon'\|O^{(\mathrm{unk})}\|^2+\epsilon'\|O^{(\mathrm{unk})}\|^{2-r}\|O^{(\mathrm{low})}\|^r \tag{D84}$$

with probability at least $1-\delta$, as claimed in Eq. (D3).

Finally, by noting that $|S|=\mathcal{O}(n^k)$ and $k=\log_{1.5}(1/\epsilon)$, and recalling the definition of $\tilde{\epsilon}$ in Eq .(D41) on the right-hand side of Eq. (D68), we have

$$\begin{aligned}\frac{\log(1/\delta)}{\epsilon'}+\frac{\log(|S|/\delta)}{\tilde{\epsilon}^2}&=\log\left(\frac{n}{\delta}\right)\left(\frac{1}{\epsilon'}\right)^{k+1}2^{\mathcal{O}(k\log k)}\\&=\log\left(\frac{n}{\delta}\right)2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}.\end{aligned} \tag{D85}$$

So it suffices to have

$$N_{\mathrm{val}}=\log\left(\frac{n}{\delta}\right)2^{\Omega\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}. \tag{D86}$$

Furthermore, by noting that $R=\log_2\lceil 1/\tilde{\epsilon}\rceil=\mathcal{O}(k\log(\epsilon')+k\log^2 k)$ in Eq. (D81), we see that it suffices to have

$$N_{\mathrm{val}}=\Omega\left(\frac{\log\log(\epsilon)+\log\log(\epsilon')+\log(1/\delta)}{(\epsilon')^2}\right). \tag{D87}$$

Recall that the full data size $N=N_{\mathrm{tr}}+N_{\mathrm{val}}$, and the quantity in Eq. (D87) is dominated by that in Eq. (D86), yielding one argument in the minimum of the sample complexity claimed in Theorem 13.

d. Establishing sample complexity on the right of Eq. (D39)

By considering the full dataset size to be

$$N=\Omega\left(\frac{\log(|S'|/\delta)}{(\tilde{\epsilon}')^2}\right), \tag{D88}$$

Hoeffding's inequality and the union bound can be used to guarantee that

$$|\hat{x}'_P-x_P|\le\|O^{(\mathrm{unk})}\|\tilde{\epsilon}',\qquad|\hat{\beta}'_P-\beta_P|\le\tilde{\epsilon}',\quad\text{for all }P\in S' \tag{D89}$$

with probability at least $1-\delta$. Using Lemma 17 on filtering the small-weight factor, we have

$$\beta_P|\hat{\alpha}'_P - \alpha_P|^2 \leq 3\|O^{(\mathrm{unk})}\|^2\tilde{\epsilon}'. \tag{D90}$$

Using Lemma 14 on the mean squared error and Corollary 13 on the low-degree approximation, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(\hat{O}'\rho) - \operatorname{tr}(O^{(\mathrm{unk})}\rho)|^2 \\
&\quad\leq (2/3)^k\|O^{(\mathrm{unk})}\|^2 + \sum_{P\in S'}\beta_P|\hat{\alpha}'_P - \alpha_P|^2 \\
&\quad\leq \|O^{(\mathrm{unk})}\|^2\frac{\epsilon}{2} + 3n^{k'}\|O^{(\mathrm{unk})}\|^2\tilde{\epsilon}'.
\end{aligned} \tag{D91}$$

From the definition of $\tilde{\epsilon}'$ in Eq. (D53), we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |\operatorname{tr}(\hat{O}'\rho) - \operatorname{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \epsilon\|O^{(\mathrm{unk})}\|^2. \tag{D92}$$

The sample complexity is

$$N = \mathcal{O}\left(\frac{\log(|S'|/\delta)}{(\tilde{\epsilon}')^2}\right) = \log(n/\delta)\, 2^{\mathcal{O}(\log(1/\epsilon)\log(n))}, \tag{D93}$$

which completes the sample complexity claimed in Theorem 13.

APPENDIX E: LEARNING QUANTUM EVOLUTIONS FROM RANDOMIZED EXPERIMENTS

We recall the following definitions pertaining to classical shadows for quantum states and quantum evolutions, based on randomized Pauli measurements and random input states.

Definition 7 (Single-qubit stabilizer state).—We define

$$\mathrm{stab}_1 \triangleq \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\} \tag{E1}$$

to be the set of single-qubit stabilizer states.

We define randomized Pauli measurements as follows.

Definition 8 (Randomized Pauli measurement).—Let $n > 0$. A randomized Pauli measurement on an n-qubit state is given by a 6^n-outcome positive operator-valued measure (POVM)

$$\mathcal{F}^{(\mathrm{Pauli})} \triangleq \left\{\frac{1}{3^n}\bigotimes_{i=1}^{n}|s_i\rangle\langle s_i|\right\}_{s_1,\ldots,s_n\in\mathrm{stab}_1}, \tag{E2}$$

which corresponds to measuring every qubit under a random Pauli basis (X, Y, Z). The outcome of $\mathcal{F}^{(\mathrm{Pauli})}$ is an n-qubit state $|\psi\rangle = \bigotimes_{i=1}^{n}|s_i\rangle$, where $|s_i\rangle \in \mathrm{stab}_1$ is a single-qubit stabilizer state.

In the following, we define the classical shadow of a quantum state based on randomized Pauli measurements. Classical shadows could also be defined based on other randomized measurements [46].

Definition 9 (Classical shadow of a quantum state).—Let $n, N > 0$. Consider an n-qubit state ρ. A size-N classical shadow $S_N(\rho)$ of quantum state ρ is a random set given by

$$S_N(\rho) \triangleq \{|\psi_\ell\rangle\}_{\ell=1}^{N}, \tag{E3}$$

where $|\psi_\ell\rangle = \bigotimes_{i=1}^{n}|s_{\ell,i}\rangle$ is the outcome of the ℓth randomized Pauli measurement on a single copy of ρ.

We can generalize classical shadows from quantum states to quantum processes by considering random product input states and randomized Pauli measurements. A similar generalization has been studied in Ref. [33].

Definition 10 (Classical shadow of a quantum process).—Consider an n-qubit CPTP map $\mathcal{E}$. A size-N classical shadow $S_N(\mathcal{E})$ of quantum evolution $\mathcal{E}$ is a random set given by

$$S_N(\mathcal{E}) \triangleq \{|\psi_\ell^{(\mathrm{in})}\rangle, |\psi_\ell^{(\mathrm{out})}\rangle\}_{\ell=1}^{N}, \tag{E4}$$

where $|\psi_\ell^{(\mathrm{in})}\rangle = \bigotimes_{i=1}^{n}|s_{\ell,i}^{(\mathrm{in})}\rangle$ is a random input state with $|s_{\ell,i}^{(\mathrm{in})}\rangle \in \mathrm{stab}_1$ sampled uniformly, and $|\psi_\ell^{(\mathrm{out})}\rangle = \bigotimes_{i=1}^{n}|s_{\ell,i}^{(\mathrm{out})}\rangle$ is the outcome of performing the randomized Pauli measurement on $\mathcal{E}(|\psi_\ell^{(\mathrm{in})}\rangle\langle\psi_\ell^{(\mathrm{in})}|)$.

After obtaining the outcome from N randomized experiments, we can design a learning algorithm that learns a model of the unknown CPTP map $\mathcal{E}$ such that, given an input state ρ and an observable O, the algorithm could predict $\operatorname{tr}(O\mathcal{E}(\rho))$. The rigorous guarantee is given in the following theorem.

Theorem 14 (Learning to predict a quantum evolution).—Let $n, \epsilon, \epsilon', \delta > 0$. Consider any unknown n-qubit CPTP map $\mathcal{E}$, and a classical shadow $S_N(\mathcal{E})$ of $\mathcal{E}$ obtained by N randomized experiments with

$$\begin{aligned}
N = \log\left(\frac{n}{\delta}\right)\min(&2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}, \\
&\times 2^{\mathcal{O}[\log(1/\epsilon)\log(n)]}).
\end{aligned} \tag{E5}$$

With probability $\geq 1-\delta$, the algorithm learns a function h such that, for any n-qubit state distribution $\mathcal{D}$ invariant under single-qubit H and S gates, and any observable O given as a sum of few-body observables, where each qubit is acted on by $\mathcal{O}(1)$ of the few-body observables,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}} |h(\rho, O) - \operatorname{tr}(O\mathcal{E}(\rho))|^2 \\
&\quad\leq \left(\epsilon + \epsilon'\left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^{2\lceil\log_{1.5}(1/\epsilon)\rceil/[\lceil\log_{1.5}(1/\epsilon)\rceil+1]}\right)\|O\|^2.
\end{aligned} \tag{E6}$$

Here, $O^{(\mathrm{low})}$ is the low-degree approximation of O after Heisenberg evolution under $\mathcal{E}$.

The scaling given in the main text corresponds to the additional assumption that $\|O\| \leq 1$. By noting that $2\lceil \log_{1.5}(1/\epsilon)\rceil / [\lceil \log_{1.5}(1/\epsilon)\rceil + 1] \in [1, 2)$, we have

$$\begin{aligned} \left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^{2\lceil \log_{1.5}(1/\epsilon)\rceil/[\lceil \log_{1.5}(1/\epsilon)\rceil+1]} & \|O\|^2 \\ &\leq \|O^{(\mathrm{low})}\|^{2\lceil \log_{1.5}(1/\epsilon)\rceil/[\lceil \log_{1.5}(1/\epsilon)\rceil+1]} \\ &\leq \max(\|O^{(\mathrm{low})}\|^2, 1). \end{aligned} \tag{E7}$$

Theorem 1 follows by considering $\epsilon' \to 0$.

1. Learning algorithm

Recall that a size-N classical shadow $S_N(\mathcal{E})$ of the CPTP map $\mathcal{E}$ is a set given by

$$S_N(\mathcal{E}) \triangleq \left\{ |\psi_\ell^{(\mathrm{in})}\rangle = \bigotimes_{i=1}^{n} |s_{\ell,i}^{(\mathrm{in})}\rangle \,,\, |\psi_\ell^{(\mathrm{out})}\rangle = \bigotimes_{i=1}^{n} |s_{\ell,i}^{(\mathrm{out})}\rangle \right\}_{\ell=1}^{N}. \tag{E8}$$

Given an observable O that can be written as a sum of κ-qubit observables, where each qubit is acted on by at most d of the κ-qubit observables with $\kappa, d = \mathcal{O}(1)$, we have

$$O = \sum_{Q \in \{I,X,Y,Z\}^{\otimes n} :\, |Q| \leq \kappa} a_Q Q, \tag{E9}$$

where $\sum_{Q:\, |Q| \leq \kappa} \mathbb{1}[a_Q \neq 0] = \mathcal{O}(n)$. The algorithm creates a dataset,

$$\begin{aligned} \Bigg\{ \rho_\ell = |\psi_\ell^{(\mathrm{in})}\rangle\langle\psi_\ell^{(\mathrm{in})}|,\, & y_\ell(O) \\ &= \sum_{Q:\, |Q| \leq \kappa} a_Q \operatorname{tr}\left(Q \bigotimes_{i=1}^{n} (3|s_{\ell,i}^{(\mathrm{out})}\rangle\langle s_{\ell,i}^{(\mathrm{out})}| - I) \right) \Bigg\}_{\ell=1}^{N} \end{aligned} \tag{E10}$$

from the classical shadow $S_N(\mathcal{E})$, which requires $\mathcal{O}(nN)$ computational time. We also define the parameter

$$\eta \triangleq \sum_{Q:\, |Q| \leq \kappa} |a_Q| = \|O\|_{\mathrm{Pauli},1} \tag{E11}$$

based on the given observable O.

The sample complexity in Eq. (E5) is the minimum of two arguments. Each of the two corresponds to a hyperparameter setting for k and $\tilde{\epsilon}$. Let $C(k)$ be the function from Corollary 11 and $C(k, d)$ be the function from Corollary 12. The first hyperparameter setting considers

$$\begin{aligned} k = \lceil \log_{1.5}(1/\epsilon)\rceil, \qquad \tilde{\epsilon} &= \left(\frac{\epsilon'}{6 \cdot 2^k}\right)^{k+1} \\ &\times \left(\frac{C(\kappa, d)}{3}\right)^2 \left(\frac{C(k)}{3}\right)^{2k}. \end{aligned} \tag{E12}$$

The second hyperparameter setting considers

$$k = \lceil \log_{1.5}(2/\epsilon)\rceil, \qquad \tilde{\epsilon} = \frac{\epsilon}{9 \cdot 2^{k+1} n^k} \left(\frac{C(\kappa, d)}{3}\right)^2. \tag{E13}$$

For every Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$ with $|P| \leq k$, the algorithm computes

$$\hat{x}_P(O) = \frac{1}{N} \sum_{\ell=1}^{N} \operatorname{tr}(P \rho_\ell) y_\ell(O), \tag{E14}$$

$$\hat{\beta}_P = \left(\frac{1}{3}\right)^{|P|}, \tag{E15}$$

$$\hat{\alpha}_P(O) = \begin{cases} 0, & \hat{\beta}_P \leq 2\tilde{\epsilon}, \\ 0, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P(O)|/\hat{\beta}_P^{1/2} \leq 2\eta\sqrt{\tilde{\epsilon}}, \\ \hat{x}_P(O)/\hat{\beta}_P, & \hat{\beta}_P > 2\tilde{\epsilon},\ |\hat{x}_P(O)|/\hat{\beta}_P^{1/2} > 2\eta\sqrt{\tilde{\epsilon}}, \end{cases} \tag{E16}$$

which requires $\mathcal{O}(kN)$ time per Pauli observable P. Finally, given an n-qubit state ρ, the algorithm outputs

$$h(\rho, O) \triangleq \sum_{P:\, |P| \leq k} \hat{\alpha}_P(O) \operatorname{tr}(P\rho), \tag{E17}$$

which uses a computational time of $\mathcal{O}(n^k)$.

2. Rigorous performance guarantee

In this section, we prove that the learning algorithm presented in the last section satisfies Theorem 14. The proof uses the tools presented in Appendix D 2 and is similar to the proof of Theorem 13.

a. Definitions

For a given observable that is a sum of κ-qubit observables, where $\kappa = \mathcal{O}(1)$ and each qubit is acted on by

$d = \mathcal{O}(1)$ of the κ-qubit observables, we can write

$$O = \sum_{Q \in \{I,X,Y,Z\}^{\otimes n}:\, |Q| \leq \kappa} a_Q Q. \tag{E18}$$

We define a few variables based on O as follows. We consider the unknown observable to be

$$O^{(\mathrm{unk})} \triangleq \mathcal{E}^{\dagger}(O) \triangleq \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \alpha_P(O) P, \tag{E19}$$

and the low-degree approximation of $O^{(\mathrm{unk})}$ to be

$$O^{(\mathrm{low})} \triangleq \sum_{P \in \{I,X,Y,Z\}^{\otimes n}:\, |P| \leq k} \alpha_P(O) P. \tag{E20}$$

Then, for all Pauli observables $P \in \{I,X,Y,Z\}^{\otimes n}$, we define

$$x_P(O) \triangleq \left(\frac{1}{3}\right)^{|P|} \alpha_P(O), \qquad \beta_P \triangleq \left(\frac{1}{3}\right)^{|P|}. \tag{E21}$$

We also define the standard n-qubit input state distribution $\mathcal{D}^0$ to be the uniform distribution over the tensor product of n single-qubit stabilizer states. A nice property of $\mathcal{D}^0$ is that, for any state ρ in the support of $\mathcal{D}^0$, the nonidentity purity for a subsystem A of size L is

$$\gamma^*(\rho_A) = \frac{1}{2^L}. \tag{E22}$$

Using this property and Lemma 16 on extracting Pauli coefficients, we have the identities

$$x_P(O) = \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} \operatorname{tr}(P\rho) \operatorname{tr}(\mathcal{E}^{\dagger}(O)\rho), \qquad \beta_P = \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} \operatorname{tr}(P\rho)^2$$
$$= \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} [\gamma^*(\rho_{\mathsf{dom}(P)})] \left(\frac{2}{3}\right)^{|P|}. \tag{E23}$$

We are now ready to prove Theorem 14.

b. Prediction error under the standard distribution $\mathcal{D}^0$ (first set of hyperparameters)

We begin the proof by considering the first set of hyperparameters $k, \tilde{\epsilon}$ as given in Eq. (E12). For a Pauli observable $Q \in \{I,X,Y,Z\}^{\otimes n}$ with $|Q| \leq \kappa = \mathcal{O}(1)$, we consider the random variable

$$\hat{x}_P(Q) = \frac{1}{N} \sum_{\ell=1}^{N} \operatorname{tr}(P\rho_\ell) y_\ell(Q) = \frac{1}{N} \sum_{\ell=1}^{N} \operatorname{tr}(P\rho_\ell) \operatorname{tr}$$
$$\times \left(Q \bigotimes_{i=1}^{n} (3|s_{\ell,i}^{(\mathrm{out})}\rangle\langle s_{\ell,i}^{(\mathrm{out})}| - I)\right). \tag{E24}$$

Because $|Q| = \mathcal{O}(1)$, we have $|\operatorname{tr}(Q \bigotimes_{i=1}^{n} (3|s_{\ell,i}^{(\mathrm{out})}\rangle\langle s_{\ell,i}^{(\mathrm{out})}| - I))| = \mathcal{O}(1)$ with probability 1. By considering the size of the classical shadow $S_N(\mathcal{E})$ to be

$$N = \Omega\left(\frac{\log(n^{k+\kappa}/\delta)}{\tilde{\epsilon}^2}\right), \tag{E25}$$

we can utilize Hoeffding's inequality and the union bound to guarantee that

$$|\hat{x}_P(Q) - x_P(Q)| \leq \tilde{\epsilon} \quad \text{for all } P, Q \in \{I,X,Y,Z\}^{\otimes n}, |P|$$
$$\leq k,\ |Q| \leq \kappa \tag{E26}$$

with probability at least $1 - \delta$. In the following proof, we condition on the above event.

Using the triangle inequality, we have

$$|\hat{x}_P(O) - x_P(O)| \leq \|O\|_{\mathrm{Pauli},1} \tilde{\epsilon} = \eta\tilde{\epsilon},$$
$$|\hat{\beta}_P - \beta_P| = 0, \quad \text{for all } P\colon |P| \leq k. \tag{E27}$$

The norm inequality given in Corollary 11 shows that

$$\sum_{P\colon |P| \leq k} |\alpha_P(O)|^r \leq \left(\frac{3}{C(k)}\right)^r \|O^{(\mathrm{low})}\|^r \tag{E28}$$

for the constant $C(k)$ defined in Eq. (5).

The filtering lemma given in Lemma 18 shows that

$$\sum_{P\colon |P| \leq k} \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} [\gamma^*(\rho_{\mathsf{dom}(P)})] \left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2$$
$$\leq 6\eta^{2-r} \left(\frac{3}{C(k)}\right)^r \|O^{(\mathrm{low})}\|^r \tilde{\epsilon}^{1-r/2}. \tag{E29}$$

From the norm inequality and constant $C(k,d)$ given in Corollary 12, we have

$$\eta = \|O\|_{\mathrm{Pauli},1} \leq \frac{3}{C(\kappa,d)} \|O\|. \tag{E30}$$

Combined with the definition of $\tilde{\epsilon}$ given in Eq. (E12), we have

$$\sum_{P\colon |P| \leq k} \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} [\gamma^*(\rho_{\mathsf{dom}(P)})] \left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2$$
$$\leq \left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^r \frac{\epsilon'}{2^k} \|O\|^2. \tag{E31}$$

Using Lemma 14 on the mean squared error and Corollary 13 on the low-degree approximation, we have

$$\mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} |h(\rho, O) - \operatorname{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \underbrace{(2/3)^k \|O^{(\mathrm{unk})}\|^2}_{\leq \|O^{(\mathrm{unk})}\|^2 \epsilon}$$
$$+ \sum_{P\colon |P| \leq k} \mathop{\mathbb{E}}_{\rho \sim \mathcal{D}^0} [\gamma^*(\rho_{\mathsf{dom}(P)})] \left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2. \tag{E32}$$

Using the definition of $O^{(\mathrm{unk})}$, we have $O^{(\mathrm{unk})} = \mathcal{E}^\dagger(O)$ and $\|O^{(\mathrm{unk})}\| \leq \|O\|$. Hence,

$$\underset{\rho\sim\mathcal{D}^0}{\mathbb{E}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2 \leq \left(\epsilon + \frac{\epsilon'}{2^k}\left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^r\right)\|O\|^2, \tag{E33}$$

which establishes a prediction error bound for distribution $\mathcal{D}^0$.

c. Prediction error under the general distribution $\mathcal{D}$ (first set of hyperparameters)

We now consider an arbitrary n-qubit state distribution $\mathcal{D}$ invariant under single-qubit H and S gates. Using Lemma 14 on the mean squared error and Corollary 13 on the low-degree approximation, we have

$$\underset{\rho\sim\mathcal{D}}{\mathbb{E}} |h(\rho, O) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \epsilon\|O\|^2 + \sum_{P:\,|P|\leq k} \underset{\rho\sim\mathcal{D}}{\mathbb{E}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2. \tag{E34}$$

Recall that $\gamma^*(\rho_{\mathsf{dom}(P)}) \leq 1$; hence,

$$\underset{\rho\sim\mathcal{D}}{\mathbb{E}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|} \leq 2^k\left(\frac{1}{3}\right)^{|P|} \quad \text{for all } P \in \{I, X, Y, Z\}^{\otimes n},\ |P| \leq k. \tag{E35}$$

Furthermore, we have $\mathbb{E}_{\rho\sim\mathcal{D}_0}[\gamma^*(\rho_{\mathsf{dom}(P)})](2/3)^{|P|} = (1/3)^{|P|}$. Together, we have

$$\underset{\rho\sim\mathcal{D}}{\mathbb{E}} |h(\rho, O) - \mathrm{tr}(O^{(\mathrm{unk})}\rho)|^2 \leq \epsilon\|O\|^2 + 2^k \times \sum_{P:\,|P|\leq k} \underset{\rho\sim\mathcal{D}^0}{\mathbb{E}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2. \tag{E36}$$

Combining the above with Eq. (E31), we have

$$\underset{\rho\sim\mathcal{D}}{\mathbb{E}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2 \leq \left(\epsilon + \epsilon'\left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^r\right)\|O\|^2, \tag{E37}$$

which is the prediction error under distribution $\mathcal{D}$.

d. Putting everything together (first set of hyperparameters)

From Eq. (E12), we have set the parameter $\tilde{\epsilon}$ to be

$$\tilde{\epsilon} = \left(\frac{\epsilon'}{6}\right)^{k+1}\left(\frac{C(\kappa, d)}{3}\right)^2\left(\frac{C(k)}{3}\right)^{2k}. \tag{E38}$$

Furthermore, given the classical shadow $S_N(\mathcal{E})$ of size

$$N = \mathcal{O}\left(\frac{\log(n^{k+\kappa}/\delta)}{\tilde{\epsilon}^2}\right) = \log\left(\frac{n}{\delta}\right) \times 2^{\mathcal{O}\{\log(1/\epsilon)[\log\log(1/\epsilon)+\log(1/\epsilon')]\}}, \tag{E39}$$

we can guarantee that, with probability at least $1 - \delta$, the following holds. For any observable O that is a sum of κ-qubit observables, where $\kappa = \mathcal{O}(1)$ and each qubit is acted on by $d = \mathcal{O}(1)$ of the κ-qubit observables, and any n-qubit state distribution $\mathcal{D}$ invariant under single-qubit H and S gates, we have

$$\underset{\rho\sim\mathcal{D}}{\mathbb{E}} |h(\rho, O) - \mathrm{tr}(O\mathcal{E}(\rho))|^2 \leq \left(\epsilon + \epsilon'\left[\frac{\|O^{(\mathrm{low})}\|}{\|O\|}\right]^r\right)\|O\|^2. \tag{E40}$$

This establishes one of the arguments for the sample complexity stated in Theorem 14.

e. Prediction error under the standard distribution $\mathcal{D}^0$ (second set of hyperparameters)

In the following proof, we consider the second set of hyperparameters $k, \tilde{\epsilon}$ as given in Eq. (E13). By considering the size of the classical shadow $S_N(\mathcal{E})$ to be

$$N = \Omega\left(\frac{\log(n^{k+\kappa}/\delta)}{\tilde{\epsilon}^2}\right), \tag{E41}$$

we can utilize Hoeffding's inequality and the union bound to guarantee that

$$|\hat{x}_P(Q) - x_P(Q)| \leq \tilde{\epsilon} \quad \text{for all } P, Q \in \{I, X, Y, Z\}^{\otimes n},\ |P| \leq k,\ |Q| \leq \kappa \tag{E42}$$

with probability at least $1 - \delta$. In the following proof, we condition on the above event. Using the triangle inequality, we have

$$|\hat{x}_P(O) - x_P(O)| \leq \|O\|_{\mathrm{Pauli},1}\tilde{\epsilon} = \eta\tilde{\epsilon}, \quad |\hat{\beta}_P - \beta_P| = 0, \quad \text{for all } P\colon |P| \leq k. \tag{E43}$$

The filtering lemma given in Lemma 18 shows that

$$\sum_{P:\,|P|\leq k} \underset{\rho\sim\mathcal{D}^0}{\mathbb{E}}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|} |\hat{\alpha}_P(O) - \alpha_P(O)|^2 \leq 9\eta^2\tilde{\epsilon}^2. \tag{E44}$$

From the norm inequality and function $C(k,d)$ given in Corollary 12, we have

$$\eta = \|O\|_{\text{Pauli},1} \leq \frac{3}{C(\kappa,d)}\|O\|. \tag{E45}$$

Combined with the definition of $\tilde{\epsilon}$ given in Eq. (E13), we have

$$\sum_{P:\,|P|\leq k} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(O) - \alpha_P(O)|^2 \leq \frac{\epsilon}{2^{k+1}}\|O\|^2. \tag{E46}$$

Using Lemma 14 on the mean squared error and Corollary 13 on the low-degree approximation, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}|h(\rho,O) - \operatorname{tr}(O^{(\text{unk})}\rho)|^2 &\leq (2/3)^k\|O^{(\text{unk})}\|^2 \\ &+ \sum_{P:\,|P|\leq k} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(O) - \alpha_P(O)|^2 \\ &\leq \frac{\epsilon}{2}\|O^{(\text{unk})}\|^2 + \sum_{P:\,|P|\leq k} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|} \\ &|\hat{\alpha}_P(O) - \alpha_P(O)|^2.\end{aligned} \tag{E47}$$

Using the definition of $O^{(\text{unk})}$, we have $O^{(\text{unk})} = \mathcal{E}^\dagger(O)$ and $\|O^{(\text{unk})}\| \leq \|O\|$. Hence,

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}|h(\rho,O) - \operatorname{tr}(O\mathcal{E}(\rho))|^2 \leq \frac{1}{2}\left(\epsilon + \frac{\epsilon}{2^k}\right)\|O\|^2, \tag{E48}$$

which establishes a prediction error bound for distribution $\mathcal{D}^0$.

f. Prediction error under the general distribution $\mathcal{D}$ (second set of hyperparameters)

We now consider an arbitrary n-qubit state distribution $\mathcal{D}$ invariant under single-qubit H and S gates. Using Lemma 14 on the mean squared error, Corollary 13 on the low-degree approximation, $k = \lceil\log_{1.5}(2/\epsilon)\rceil$, the fact that $\gamma^*(\rho_{\mathsf{dom}(P)}) \leq 1$, and $\mathbb{E}_{\rho\sim\mathcal{D}_0}[\gamma^*(\rho_{\mathsf{dom}(P)})](2/3)^{|P|} = (1/3)^{|P|}$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho,O) - \operatorname{tr}(O^{(\text{unk})}\rho)|^2 &\leq \frac{\epsilon}{2}\|O\|^2 + 2^k \\ &\times \sum_{P:\,|P|\leq k} \mathop{\mathbb{E}}_{\rho\sim\mathcal{D}^0}[\gamma^*(\rho_{\mathsf{dom}(P)})]\left(\frac{2}{3}\right)^{|P|}|\hat{\alpha}_P(O) - \alpha_P(O)|^2.\end{aligned} \tag{E49}$$

Combining the above with Eq. (E46), we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho,O) - \operatorname{tr}(O\mathcal{E}(\rho))|^2 \leq \epsilon\|O\|^2, \tag{E50}$$

which is the prediction error under distribution $\mathcal{D}$.

g. Putting everything together (second set of hyperparameters)

From Eq. (E13), we have set the parameter $\tilde{\epsilon}$ to be

$$\tilde{\epsilon} = \frac{\epsilon}{9\cdot 2^{k+1}n^k}\left(\frac{C(\kappa,d)}{3}\right)^2. \tag{E51}$$

Furthermore, given the classical shadow $S_N(\mathcal{E})$ of size

$$N = \mathcal{O}\left(\frac{\log(n^{k+\kappa}/\delta)}{\tilde{\epsilon}^2}\right) = \log\left(\frac{n}{\delta}\right)2^{\mathcal{O}(\log(1/\epsilon)\log(n))}, \tag{E52}$$

we can guarantee that, with probability at least $1-\delta$, the following holds. For any observable O that is a sum of κ-qubit observables, where $\kappa = \mathcal{O}(1)$ and each qubit is acted on by $d = \mathcal{O}(1)$ of the κ-qubit observables, and any n-qubit state distribution $\mathcal{D}$ invariant under single-qubit H and S gates, we have

$$\mathop{\mathbb{E}}_{\rho\sim\mathcal{D}}|h(\rho,O) - \operatorname{tr}(O\mathcal{E}(\rho))|^2 \leq \epsilon\|O\|^2. \tag{E53}$$

This completes the proof of Theorem 14.

APPENDIX F: NUMERICAL DETAILS

In the numerical experiments, we consider two classes of Hamiltonians:

$$H = \frac{1}{4}\sum_i (X_iX_{i+1} + Y_iY_{i+1}) + \frac{1}{2}\sum_i h_iZ_i \quad (XY\text{model}), \tag{F1}$$

$$H = \frac{1}{2}\sum_i X_iX_{i+1} + \frac{1}{2}\sum_i h_iZ_i \quad (\text{Ising model}). \tag{F2}$$

Here $h_i = 0.5$ for the homogeneous Z field, and h_i is sampled uniformly at random from $[-5,5]$ for the disordered Z field. We solve for the time-evolved properties using the Jordan-Wigner transform to map the spin chains to a free-fermion model and the technique described in Ref. [83] to solve the free-fermion model.

We consider the training set to be a collection of N random product states $|\psi_\ell\rangle$, $\ell = 1,\ldots,N$, and their associated measured properties y_ℓ corresponding to measuring an observable O after evolving under $U(t) = \exp(-itH)$. The measured properties are averaged over 500 measurements. Hence, y_ℓ is a noisy estimate of the true expectation value $\operatorname{tr}(OU(t)|\psi_\ell\rangle\langle\psi_\ell|U(t)^\dagger)$. We consider essentially the same ML algorithm as described in Sec. III A, but utilize a more sophisticated approach to enforce sparsity in $\hat{\alpha}_P$. We

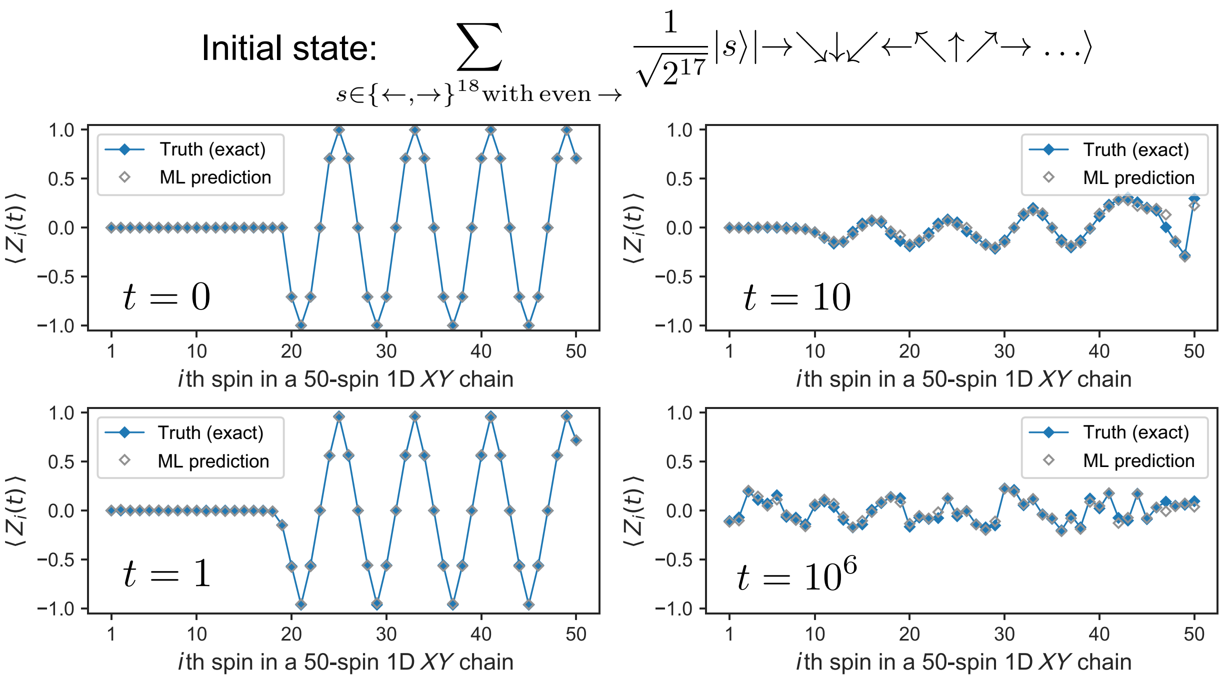

FIG. 5. Visualization of the ML model's prediction for a highly entangled initial state $\rho = |\psi\rangle\langle\psi|$. We consider the expected value of $Z_i(t) = e^{itH} Z_i e^{-itH}$, where H corresponds to the 1D 50-spin XY chain with a homogeneous Z field. The initial state $|\psi\rangle$ has a GHZ-like entanglement over the first 18-spin chain and is a product state with spins rotating clockwise over the latter 32-spin chain. To prepare $|\psi\rangle$ with 1D circuits, a depth of at least $\Omega(n)$ is required. Even though the ML model is trained only on random product states (a total of $N = 10\,000$), it still performs accurately in predicting the highly entangled state over a wide range of evolution time t.

also consider α_P for Pauli operator P that is geometrically local. For ease of analysis, we consider a simple strategy of setting small values to zero. The standard approach that is often used in practice is LASSO [73].

In the numerical experiments, we perform a simple grid search for the two hyperparameters using twofold cross-validation on the training set:

$$k = 1, 2, 3, 4, \qquad a = 2^{-15}, 2^{-14}, 2^{-13}, \ldots, 2^{-4}, 2^{-3}. \tag{F3}$$

Here k corresponds to the maximum number of qubits that the Pauli operators P act on, and a is a hyperparameter corresponding to the strength of the ℓ_1-regularization term in LASSO. In particular, the optimization problem of LASSO is given by

$$\min_{\hat{\alpha}_P} \frac{1}{2N} \sum_{\ell=1}^{N} \left| y_\ell - \sum_{P:\,|P|\leq k} \hat{\alpha}_P \operatorname{tr}(P|\psi_\ell\rangle\langle\psi_\ell|) \right|^2 + a \sum_{P:\,|P|\leq k} |\hat{\alpha}_P|, \tag{F4}$$

where $|P|$ is the number of qubits that the Pauli observable P acts nontrivially on. We then use the values $\hat{\alpha}_P$ found by the above optimization to form a succinct approximate model

$$\sum_{P:\,|P|\leq k} \hat{\alpha}_P P \tag{F5}$$

of the time-evolved observable $O(t) = U(t)^\dagger O U(t)$. Given a new initial state ρ, we predict the time-evolved property $\operatorname{tr}(O(t)\rho) = \operatorname{tr}(OU(t)\rho U(t)^\dagger)$ using

$$\sum_{P:\,|P|\leq k} \hat{\alpha}_P \operatorname{tr}(P\rho). \tag{F6}$$

In addition to the figures given in the main text, Fig. 5 shows another example for predicting a highly entangled initial state. Even though the ML model is trained with random product states, it still performs very well on a structured entangled initial state.

[1] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd, Quantum machine learning, Nature **549**, 195 (2017).

[2] Maria Schuld and Nathan Killoran, Quantum machine learning in feature Hilbert spaces, Phys. Rev. Lett. **122**, 040504 (2019).

[3] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow,

and Jay M. Gambetta, Supervised learning with quantum-enhanced feature spaces, Nature **567**, 209 (2019).

[4] Matthias C. Caro, Hsin-Yuan Huang, Marco Cerezo, Kunal Sharma, Andrew Sornborger, Lukasz Cincio, and Patrick J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 1 (2022).

[5] Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer, Classical surrogates for quantum learning models. arXiv preprint arXiv:2206.11740, (2022).

[6] Jarrod R. McClean, Sergio Boixo, Vadim N. Smelyanskiy, Ryan Babbush, and Hartmut Neven, Barren plateaus in quantum neural network training landscapes, Nat. Commun. **9**, 1 (2018).

[7] Matthias C. Caro, Hsin-Yuan Huang, Nicholas Ezzell, Joe Gibbs, Andrew T. Sornborger, Lukasz Cincio, Patrick J. Coles, and Zoë Holmes, Out-of-distribution generalization for learning quantum dynamics. arXiv preprint arXiv:2204.10268, (2022).

[8] Hsin-Yuan Huang, Michael Broughton, Jordan Cotler, Sitan Chen, Jerry Li, Masoud Mohseni, Hartmut Neven, Ryan Babbush, Richard Kueng, John Preskill, and Jarrod R. McClean, Quantum advantage in learning from experiments, Science **376**, 1182 (2022).

[9] Edward Farhi and Hartmut Neven, Classification with quantum neural networks on near term processors. arXiv preprint arXiv:1802.06002, (2018).

[10] Srinivasan Arunachalam and Ronald de Wolf, Guest column: A survey of quantum learning theory, ACM SIGACT News **48**, 41 (2017).

[11] Joe Gibbs, Zoë Holmes, Matthias C. Caro, Nicholas Ezzell, Hsin-Yuan Huang, Lukasz Cincio, Andrew T. Sornborger, and Patrick J. Coles, Dynamical simulation via quantum machine learning with provable generalization. arXiv preprint arXiv:2204.10269, (2022).

[12] Cristina Cirstoiu, Zoe Holmes, Joseph Iosue, Lukasz Cincio, Patrick J. Coles, and Andrew Sornborger, Variational fast forwarding for quantum simulation beyond the coherence time, Npj Quantum Inf. **6**, 1 (2020).

[13] Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J. Love, Alán Aspuru-Guzik, and Jeremy L. O'brien, A variational eigenvalue solver on a photonic quantum processor, Nat. Commun. **5**, 4213 (2014).

[14] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M. Chow, and Jay M. Gambetta, Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets, Nature **549**, 242 (2017).

[15] Christian Kokail, Christine Maier, Rick van Bijnen, Tiff Brydges, Manoj K. Joshi, Petar Jurcevic, Christine A. Muschik, Pietro Silvi, Rainer Blatt, Christian F. Roos, and P. Zoller, Self-verifying variational quantum simulation of lattice models, Nature **569**, 355 (2019).

[16] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Cole, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[17] Harper R. Grimsley, Sophia E. Economou, Edwin Barnes, and Nicholas J. Mayhall, An adaptive variational algorithm for exact molecular simulations on a quantum computer, Nat. Commun. **10**, 1 (2019).

[18] Giuseppe Carleo and Matthias Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[19] Or Sharir, Yoav Levine, Noam Wies, Giuseppe Carleo, and Amnon Shashua, Deep autoregressive models for the efficient variational simulation of many-body quantum systems, Phys. Rev. Lett. **124**, 020503 (2020).

[20] Evert P. L. Van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).

[21] Zhenpeng Zhou, Xiaocheng Li, and Richard N. Zare, Optimizing chemical reactions with deep reinforcement learning, ACS Cent. Sci. **3**, 1337 (2017).

[22] Juan Carrasquilla and Roger G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[23] Robert G. Parr, in *Horizons of Qquantum Chemistry*, edited by K. Fukui and B. Pullman (Springer, Dordrecht, Netherlands, 1980), p. 5.

[24] Richard Car and Mark Parrinello, Unified approach for molecular dynamics and density-functional theory, Phys. Rev. Lett. **55**, 2471 (1985).

[25] Axel D. Becke, A new mixing of Hartree–Fock and local density-functional theories, J. Chem. Phys. **98**, 1372 (1993).

[26] Steven R. White, Density-matrix algorithms for quantum renormalization groups, Phys. Rev. B **48**, 10345 (1993).

[27] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl, Neural message passing for quantum chemistry. arXiv preprint arXiv:1704.01212, (2017).

[28] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V. Albert, and John Preskill, Provably efficient machine learning for quantum many-body problems. arXiv preprint arXiv:2106.12627, (2021).

[29] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean, Power of data in quantum machine learning, Nat. Commun. **12**, 1 (2021).

[30] Masoud Mohseni, Ali T. Rezakhani, and Daniel A. Lidar, Quantum-process tomography: Resource analysis of different strategies, Phys. Rev. A **77**, 032322 (2008).

[31] A. J. Scott, Optimizing quantum process tomography with unitary 2-designs, J. Phys. **A41**, 055308 (2008).

[32] Jeremy L. O'Brien, Geoff J. Pryde, Alexei Gilchrist, Daniel F. V. James, Nathan K. Langford, Timothy C. Ralph, and Andrew G. White, Quantum process tomography of a controlled-NOT gate, Phys. Rev. Lett. **93**, 080502 (2004).

[33] Ryan Levy, Di Luo, and Bryan K. Clark, Classical shadows for quantum process tomography on near-term quantum computers. arXiv preprint arXiv:2110.02965, (2021).

[34] Hsin-Yuan Huang, Steven T. Flammia, and John Preskill, Foundations for learning from noisy quantum experiments. arXiv preprint arXiv:2204.13691, (2022).

[35] Seth T. Merkel, Jay M. Gambetta, John A. Smolin, Stefano Poletto, Antonio D. Córcoles, Blake R. Johnson, Colm A. Ryan, and Matthias Steffen, Self-consistent quantum process tomography, Phys. Rev. A **87**, 062119 (2013).

[36] Robin Blume-Kohout, John King Gamble, Erik Nielsen, Kenneth Rudinger, Jonathan Mizrahi, Kevin Fortier, and Peter Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, Nat. Commun. **8**, 1 (2017).

[37] Hsin-Yuan Huang, Richard Kueng, and John Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[38] Alexey A. Melnikov, Hendrik Poulsen Nautrup, Mario Krenn, Vedran Dunjko, Markus Tiersch, Anton Zeilinger, and Hans J. Briegel, Active learning machine learns to create new quantum experiments, Proc. Natl. Acad. Sci. U.S.A. **115**, 1221 (2018).

[39] Jonathan Kunjummen, Minh C. Tran, Daniel Carney, and Jacob M. Taylor, Shadow process tomography of quantum channels. arXiv preprint arXiv:2110.03629, (2021).

[40] Kai-Min Chung and Han-Hsuan Lin, Sample efficient algorithms for learning quantum channels in PAC model and the approximate state discrimination problem. arXiv preprint arXiv:1810.10938, (2018).

[41] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell, On the Fourier tails of bounded functions over the discrete cube, Israel J. Math. **160**, 389 (2006).

[42] Boaz Barak, Ankur Moitra, Ryan O'Donnell, Prasad Raghavendra, Oded Regev, David Steurer, Luca Trevisan, Aravindan Vijayaraghavan, David Witmer, and John Wright, Beating the random assignment on constraint satisfaction problems of bounded degree. In Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*, volume 40 of *Leibniz International Proceedings in Informatics (LIPIcs)*, p. 110, (Dagstuhl, Germany, 2015). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[43] Aram W. Harrow and Ashley Montanaro, Extremal eigenvalues of local Hamiltonians, Quantum **1**, 6 (2017).

[44] Anurag Anshu, David Gosset, Karen J. Morenz Korol, and Mehdi Soleimanifar, Improved approximation algorithms for bounded-degree local Hamiltonians, Phys. Rev. Lett. **127**, 250502 (2021).

[45] Cambyse Rouzé, Melchior Wirth, and Haonan Zhang, Quantum Talagrand, KKL and Friedgut's theorems and the learnability of quantum Boolean functions. (2022).

[46] Hsin-Yuan Huang, Richard Kueng, and John Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[47] Andreas Elben, Steven T. Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller, The randomized measurement toolbox. arXiv preprint arXiv:2203.11374, (2022).

[48] Julia Kempe, Alexei Kitaev, and Oded Regev, The complexity of the local Hamiltonian problem, Siam J Comput. **35**, 1070 (2006).

[49] J. J. Sakurai and Jim Napolitano, *Modern Quantum Mechanics* (Cambridge University Press, Cambridge, UK, 2017), 2nd ed.

[50] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann, A quantum approximate optimization algorithm. arXiv preprint arXiv:1411.4028, (2014).

[51] Edward Farhi, Jeffrey Goldstone, and Sam Gutmann, A quantum approximate optimization algorithm applied to a bounded occurrence constraint problem. *arXiv: Quantum Physics*, (2014).

[52] Ojas Parekh and Kevin Thompson, Beating random assignment for approximating quantum 2-local Hamiltonian problems. arXiv preprint arXiv:2012.12347, (2020).

[53] Sean Hallgren, Eun Young Lee, and Ojas Parekh, An approximation algorithm for the max-2-local Hamiltonian problem. in *APPROX-RANDOM*, (2020).

[54] Matthew B. Hastings and Ryan O'Donnell, in *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy* (Association for Computing Machinery, New York, NY, USA, 2022), p. 776.

[55] John E. Littlewood, On bounded bilinear forms in an infinite number of variables, The Quarterly Journal of Mathematics **1**, 164 (1930).

[56] H. F. Bohnenblust and Einar Hille, On the absolute convergence of Dirichlet series, Ann. Math. **32**, 600 (1931).

[57] Scott Aaronson, Shadow tomography of quantum states. in *STOC*, (2018), p. 325.

[58] Scott Aaronson and Guy N. Rothblum, Gentle measurement of quantum states and differential privacy. in *STOC*, (2019), p. 322.

[59] Andrew Zhao, Nicholas C. Rubin, and Akimasa Miyake, Fermionic partial tomography via classical shadows, Phys. Rev. Lett. **127**, 110504 (2021).

[60] Hong-Ye Hu and Yi-Zhuang You, Hamiltonian-driven shadow tomography of quantum states. arXiv preprint arXiv:2102.10132, (2021).

[61] Dax Enshan Koh and Sabee Grewal, Classical shadows with noise. arXiv preprint arXiv:2011.11580, (2020).

[62] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia, Robust shadow estimation. arXiv preprint arXiv:2011.09636, (2020).

[63] Charles Hadfield, Sergey Bravyi, Rudy Raymond, and Antonio Mezzacapo, Measurements of quantum Hamiltonians with locally-biased classical shadows. arXiv:2006.15788, (2020).

[64] G. I. Struchalin, Ya A Zagorovskii, E. V. Kovlakov, S. S. Straupe, and S. P. Kulik, Experimental estimation of quantum state properties from classical shadows. arXiv preprint arXiv:2008.05234, (2020).

[65] Hsin-Yuan Huang, Learning quantum states from their classical shadows, Nat. Rev. Phys. **4**, 81 (2022).

[66] Bryan O'Gorman, Fermionic tomography and learning. arXiv preprint arXiv:2207.14787, (2022).

[67] Kianna Wan, William J. Huggins, Joonho Lee, and Ryan Babbush, Matchgate shadows for fermionic quantum simulation. arXiv preprint arXiv:2207.13723, (2022).

[68] Kaifeng Bu, Dax Enshan Koh, Roy J. Garcia, and Arthur Jaffe, Classical shadows with Pauli-invariant unitary ensembles. arXiv preprint arXiv:2202.03272, (2022).

[69] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li, in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, Denver, CO, USA, 2022), p. 574.

[70] Luuk Coopmans, Yuta Kikuchi, and Marcello Benedetti, Predicting gibbs state expectation values with pure thermal shadows. arXiv preprint arXiv:2206.05302, (2022).

[71] Alexandros Eskenazis and Paata Ivanisvili, in *Proceedings of the 54th Annual ACM SIGACT Symposium*

on Theory of Computing, Rome, Italy (Association for Computing Machinery, New York, NY, USA, 2022), p. 203.

[72] https://github.com/hsinyuan-huang/learning-quantum-process.

[73] Robert Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc.: Series B (Methodological) **58**, 267 (1996).

[74] Giacomo Torlai, Christopher J. Wood, Atithi Acharya, Giuseppe Carleo, Juan Carrasquilla, and Leandro Aolita, Quantum process tomography with unsupervised learning and tensor networks. arXiv preprint arXiv:2006.02424, (2020).

[75] Antonio A. Gentile, Brian Flynn, Sebastian Knauer, Nathan Wiebe, Stefano Paesani, Christopher E. Granade, John G. Rarity, Raffaele Santagati, and Anthony Laing, Learning models of quantum systems from experiments, Nat. Phys. **17**, 837 (2021).

[76] Leonardo Banchi, Edward Grant, Andrea Rocchetto, and Simone Severini, Modelling non-Markovian quantum processes with recurrent neural networks, New J. Phys. **20**, 123030 (2018).

[77] Weiyuan Gong and Scott Aaronson, Learning distributions over quantum measurement outcomes. (2022).

[78] Stephen Piddock and Ashley Montanaro, The complexity of antiferromagnetic interactions and 2D lattices. arXiv preprint arXiv:1506.04014, (2015).

[79] Uffe Haagerup, The best constants in the Khintchine inequality, Stud. Math. **70**, 231 (1981).

[80] Peter Borwein and Tamás Erdélyi, *Polynomials and Polynomial Inequalities* (Springer Science & Business Media, New York, NY, USA, 1995), Vol. 161.

[81] L. LeCam, Convergence of estimates under dimensionality restrictions, Ann. Stat. **1**, 38 (1973).

[82] Yihong Wu, Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, (2017).

[83] Elliott Lieb, Theodore Schultz, and Daniel Mattis, Two soluble models of an antiferromagnetic chain, Ann. Phys. (N. Y) **16**, 407 (1961).