Learning Quantum States and Unitaries of Bounded Gate Complexity

Haimeng Zhao,[1,2,*,**] Laura Lewis,[1,3,†,**] Ishaan Kannan,[1,‡,**] Yihui Quek,[4,5,§] Hsin-Yuan Huang,[1,3,5,¶] and Matthias C. Caro[1,6,‖]

[1] *Institute for Quantum Information and Matter, Caltech, Pasadena, California 91125, USA*
[2] *Tsinghua University, Beijing 100084, China*
[3] *Google Quantum AI, Venice, California 90291, USA*
[4] *Harvard University, 17 Oxford Street, Cambridge, Massachusetts 02138, USA*
[5] *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[6] *Freie Universität Berlin, 14195 Berlin, Germany*

(Received 23 April 2024; accepted 16 September 2024; published 16 October 2024)

While quantum state tomography is notoriously hard, most states hold little interest to practically minded tomographers. Given that states and unitaries appearing in nature are of bounded gate complexity, it is natural to ask if efficient learning becomes possible. In this work, we prove that to learn a state generated by a quantum circuit with G two-qubit gates to a small trace distance, a sample complexity scaling linearly in G is necessary and sufficient. We also prove that the optimal query complexity to learn a unitary generated by G gates to a small average-case error scales linearly in G. While sample-efficient learning can be achieved, we show that under reasonable cryptographic conjectures, the computational complexity for learning states and unitaries of gate complexity G must scale exponentially in G. We illustrate how these results establish fundamental limitations on the expressivity of quantum machine-learning models and provide new perspectives on no-free-lunch theorems in unitary learning. Together, our results answer how the complexity of learning quantum states and unitaries relate to the complexity of creating these states and unitaries.

DOI: 10.1103/PRXQuantum.5.040306

I. INTRODUCTION

A central problem in quantum physics is to characterize a quantum system by constructing a full classical description of its state or its unitary evolution based on data from experiments. These two tasks, named *quantum state tomography* [1–4] and *quantum process tomography* [5–9], are (in)famous for being ubiquitous yet highly expensive. The applications of tomography include quantum metrology [10,11], verification [12,13], benchmarking [5–8,14–17], and error mitigation [18–20]. Yet tomography provably requires exponentially many (in the system size n) copies of the unknown state [21,22] or runs of the unknown process [23]. This intuitively arises from the exponential scaling of the number of parameters needed to describe an arbitrary quantum system.

But the situation is less dire than it theoretically appears. In practice, tools for analyzing many-body systems often exploit *known structures* cleverly to predict their phenomenology or classically simulate them. Notable examples include the BCS theory for superconductivity [24], tensor networks [25,26], and neural-network [27–32] ansatze. Indeed, while *most* of the states or unitaries may have exponential gate complexity [33], such objects are also unphysical: an exponentially complex state or unitary cannot be produced in nature within a reasonable amount of time [34]. In particular, in Ref. [34] it is shown that quantum states or unitaries with bounded gate complexity are precisely those that can be produced by bounded-time evolution of time-dependent local Hamiltonians.

In this work, we study whether tomography, too, can benefit from the observation that nature can only produce

*Contact author: haimengzhao@icloud.com
†Contact author: llewis@alumni.caltech.edu
‡Contact author: ishaan@alumni.caltech.edu
§Contact author: yquek@mit.edu
¶Contact author: hsinyuan@caltech.edu
‖Contact author: matthias.caro@fu-berlin.de
**These authors contributed equally to this work.

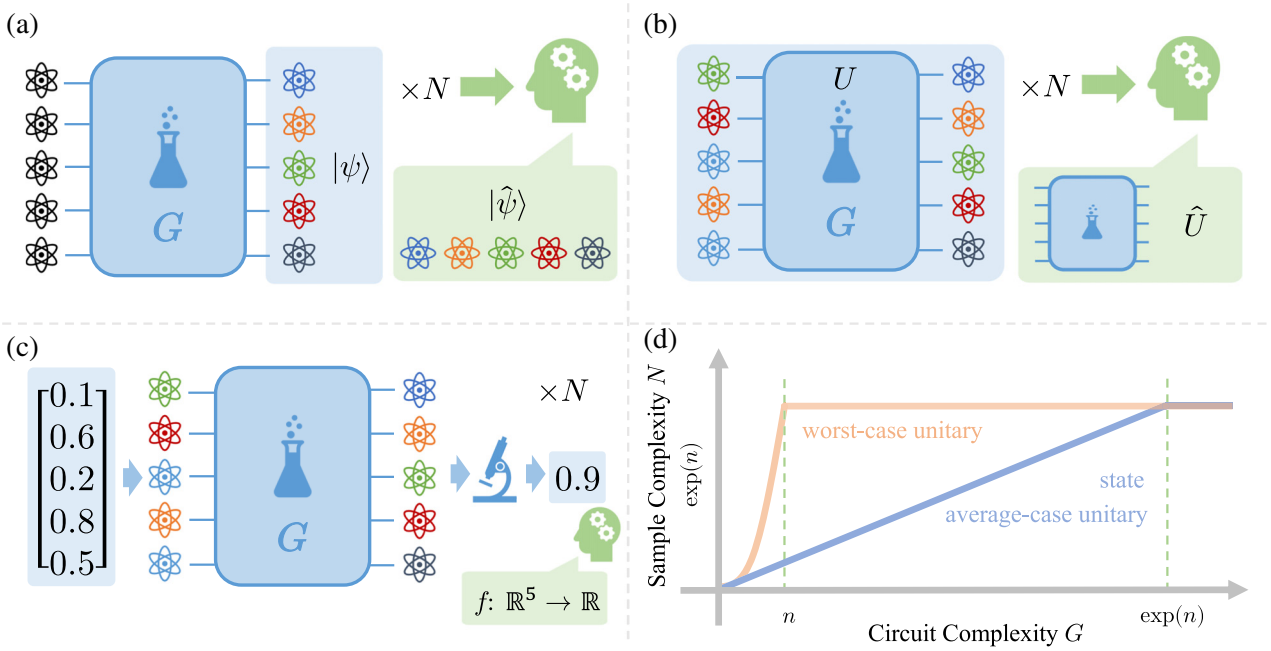

FIG. 1. (a)–(c) A schematic overview of the learning models in this work. (a) Learning quantum states with bounded circuit complexity G. (b) Learning unitaries with bounded circuit complexity G. (c) Learning classical functions from quantum experiments with bounded circuit complexity G. (d) A conceptual depiction of how the sample complexity of learning states in trace distance and unitaries in average-case distance scales linearly with circuit complexity, while that of learning unitaries in worst-case distance scales exponentially.

states and unitaries with bounded complexity. This gives rise to the following main question:

> *Can we efficiently learn states or unitaries of bounded gate complexity?*

In particular, we consider the following two tasks:

(1) Given copies (samples) of a pure quantum state $|\psi\rangle$ generated by G two-qubit gates, learn $|\psi\rangle$ to within ϵ trace distance [see Fig. 1(a)].
(2) Given uses (queries) of a unitary U composed of G two-qubit gates, learn U to within ϵ root-mean-square trace distance between output states (average-case learning) [see Fig. 1(b)].

Note that the G quantum gates can act on arbitrary pairs of qubits without any geometric locality constraint. By allowing general gates beyond discrete gate sets, this setting encompasses continuous time-dependent Hamiltonian dynamics via Trotterization [34] and thus analog quantum simulation [35]. It also includes states heavily studied in condensed matter, such as symmetry-protected topologically ordered states [36–38] and tensor-network states [39–41]. Previously, in Ref. [42] it has been shown that task (1) can be accomplished with a sample complexity of $\tilde{\mathcal{O}}(nG^2/\epsilon^4)$. In our work, we present algorithms for both of these tasks that use a number of samples or queries that are linear in the circuit complexity G up to logarithmic factors. Moreover, the sample complexity is independent of system size. Thus, for G scaling polynomially with the number of qubits, our learning procedures improve upon previous work [42] and have significantly lower sample or query complexities than required for general tomography, answering our central question in the affirmative. We also prove matching lower bounds (up to logarithmic factors), showing that our algorithms are effectively optimal. Moreover, we show that the focus on average-case learning is crucial in the case of unitaries: unitary tomography up to error ϵ in the diamond distance (a worst-case metric over input states) requires a number of queries scaling exponentially in G, establishing an exponential separation between average and worst case.

While our learning algorithms for bounded-complexity states and unitaries are efficient in terms of sample or query complexity, they are not computationally efficient. We prove that this is unavoidable. Assuming the quantum subexponential hardness of "ring learning with errors" (RingLWE) [43–48], any quantum algorithm that learns arbitrary states or unitaries with $\tilde{\mathcal{O}}(G)$ gates requires computational time scaling exponentially in G. This result highlights a significant computational-complexity limitation on learning even comparatively simple states and unitaries. This result also answers an open question in Ref. [49]. Meanwhile, we show that poly(n)-time algorithms are possible for $G = \mathcal{O}(\log n)$. Together, this

establishes a crossover in computational hardness at $G \sim \log n$, kicking in far before the sample complexity becomes exponential (at $G = \exp(n)$). This means that relatively few samples or queries already contain enough information for the learning task but it is hard to retrieve the information.

Finally, we study two variations of unitary learning that deepen our insights about the problem. The first variation utilizes classical (not quantum) descriptions of input and output pairs and explains why both learning states and unitaries display a linear-in-G sample complexity: the underlying source of complexity in learning unitaries is, in fact, the readout of input and output quantum states, rather than learning the mapping. We generalize recent quantum no-free-lunch theorems [50,51] to reach this conclusion. For the second variation, we study quantum machine-learning (QML) models. We focus on learning classical functions that map variables controlling the input states and the evolution to some experimentally observed property of the outputs [Fig. 1(c)]. Surprisingly, we find that certain well-behaved many-variable functions can in fact not (even approximately) be implemented by quantum experiments with bounded complexity. This highlights a fundamental limitation on the functional expressivity of both nature and practical QML models.

II. RESULTS

In this section, we discuss our rigorous guarantees for learning quantum states and unitaries with circuit complexity G. Our sample-complexity results are summarized in Table I and Fig. 1(d).

We also present computational-complexity results, where we establish the exponential-in-G growth of computational complexity, implying that $\log n$ gate complexity is a transition point at which learning becomes computationally inefficient. In particular, we prove that for circuit complexity $\tilde{\mathcal{O}}(G)$, any quantum algorithm for learning states in trace distance or unitaries in average-case distance must use time exponential in G, under the conjecture that RingLWE cannot be solved by a quantum computer in subexponential time. Hence, for a number G of gates that scales slightly higher than $\log n$, the learning tasks cannot be solved by any polynomial-time quantum algorithm under the same conjecture. Meanwhile, for $G = \mathcal{O}(\log n)$, both learning tasks can be solved efficiently in polynomial time.

A. Learning quantum states

We consider the task of learning quantum states of bounded circuit complexity. Let $|\psi\rangle = U|0\rangle^{\otimes n}$ be an n-qubit pure state generated by a unitary U consisting of G two-qubit gates acting on the zero state. Throughout this section, we denote $\rho \triangleq |\psi\rangle\langle\psi|$. Given N identically prepared copies of ρ, the goal is to output a classical circuit description of a quantum state $\hat{\rho}$ that is ϵ-close to ρ in trace distance: $d_{\mathrm{tr}}(\hat{\rho}, \rho) = \|\hat{\rho} - \rho\|_1/2 < \epsilon$. We establish the following theorem, which states that linear-in-G many samples (up to logarithmic factors) are both necessary and sufficient to learn the unknown quantum state $|\psi\rangle$ within a small trace distance.

Theorem 1 (State learning). Suppose that we are given N copies of an n-qubit pure state $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = U|0\rangle^{\otimes n}$ is generated by a unitary U consisting of G two-qubit gates. Then, $N = \tilde{\Theta}\left(G/\epsilon^2\right)$ copies are necessary and sufficient to learn the state within ϵ trace distance d_{tr} with high probability.

Previous work [42] has obtained a sample complexity of $\tilde{\mathcal{O}}(nG^2/\epsilon^4)$ for this task, which we show to be suboptimal. Notably, our result achieves the optimal scaling in both G and ϵ up to logarithmic factors and is independent of the system size n. Thus, we completely characterize the sample complexity, resolving an open question from Ref. [52]. We prove the upper bound in Appendix B 1, utilizing covering nets [53] and quantum hypothesis selection [54]. Our proposed algorithm first creates a covering net over the space of all unitaries consisting of G two-qubit gates. This can easily be transformed into a covering net over the space of all quantum states generated by G two-qubit gates by applying each element of the unitary covering net to the zero state. Thus, any quantum state generated by G two-qubit gates is close (in trace distance) to some element of the covering net. We can then apply quantum hypothesis selection [54] to the covering net, which allows us to identify the element in the covering net that is close to the unknown target state $|\psi\rangle$ and achieve the optimal ϵ dependence. We also note that our algorithm for learning

TABLE I. The sample complexity of learning n-qubit states and unitaries with circuit complexity G. The learning accuracy ϵ is measured in trace distance for states, the root-mean-square trace distance for average-case unitary learning, and the diamond distance for the worst case. Here, $C > 0$ is some universal constant. Throughout the paper, $\tilde{\mathcal{O}}, \tilde{\Theta}$, and $\tilde{\Omega}$ denote that we are suppressing nonleading logarithmic factors.

Sample complexity	State	Unitary (average case)	Unitary (worst case)
Upper bound	$\tilde{\mathcal{O}}\left(G/\epsilon^2\right)$	$\tilde{\mathcal{O}}\left(G\min\left\{1/\epsilon^2, \sqrt{2^n}/\epsilon\right\}\right)$	$\tilde{\mathcal{O}}\left(2^n G/\epsilon\right)$
Lower bound	$\tilde{\Omega}\left(G/\epsilon^2\right)$	$\Omega\left(G/\epsilon\right)$	$\Omega\left(2^{\min\{G/(2C), n/2\}}/\epsilon\right)$

quantum states does not require knowledge of or access to the unitary U that generates the unknown state $|\psi\rangle$. Only the condition that some unitary U consisting of G gates generates $|\psi\rangle$ is needed. The lower bound is proven in Appendix B 2 by using an information-theoretic argument via reduction to distinguishing a packing net over G-gate states [21].

Our algorithm to learn the unknown quantum state $|\psi\rangle$ is computationally inefficient, as it requires a search over a covering net the cardinality of which is exponential in G. We show that for circuits of size $\tilde{\mathcal{O}}(G)$, any quantum algorithm that can learn $|\psi\rangle$ to within a small trace distance given access to copies of this state must use time exponential in G, under commonly believed cryptographic assumptions [43–48]. Meanwhile, the learning task is computationally efficiently solvable for $G = \mathcal{O}(\log n)$ via junta learning [55] and standard tomography methods. This implies a transition point of computational efficiency at $\log n$ circuit complexity. Previous work [56,57] has arrived at similar hardness results for polynomial circuit complexity but our detailed analysis allows us to sharpen the computational lower bound and obtain this transition point. Computational-complexity lower bounds for distribution learning that are similar in spirit are also proved in Ref. [58].

Theorem 2 (State-learning computational complexity). Suppose that we are given N copies of an unknown n-qubit pure state $|\psi\rangle = U|0\rangle^{\otimes n}$ generated by an arbitrary unknown unitary U consisting of $\tilde{\mathcal{O}}(G)$ two-qubit gates. Suppose that RingLWE cannot be solved by a quantum computer in subexponential time. Then, any quantum algorithm that learns the state to within ϵ trace distance d_{tr} must use $\exp(\Omega(\min\{G, n\}))$ time. Meanwhile, for $G = \mathcal{O}(\log n)$, the learning task can be solved in polynomial time.

B. Learning quantum unitaries

For learning unitaries, a natural distance metric analogous to the trace distance for states is the diamond distance $d_\diamond(U, V) = \max_\rho \|(U \otimes I)\rho(U \otimes I)^\dagger - (V \otimes I)\rho(V \otimes I)^\dagger\|_1$, where ρ is over any arbitrarily extended Hilbert space. It characterizes the optimal success probability for discriminating between two unitary channels. Moreover, it can be reinterpreted in terms of the largest distance between $U|\psi\rangle$ and $V|\psi\rangle$ over all input states $|\psi\rangle$ and thus represents the error we make in the worst case over input states. We find that in this worst-case learning task, a number of queries exponential in G are necessary to learn the unitary.

Theorem 3 (Worst-case unitary learning). To learn an n-qubit unitary composed of G two-qubit gates to accuracy ϵ in the diamond distance $d_\diamond$ with high probability, any quantum algorithm must use at least $\Omega(2^{\min\{G/(2C), n/2\}}/\epsilon)$ queries to the unknown unitary, where $C > 0$ is a universal constant. Meanwhile, there exists such an algorithm using $\tilde{\mathcal{O}}(2^n G/\epsilon)$ queries.

The complete proof is given in Appendix C 1 and the proof of the lower bound relies on the adversary method [59–62]. We construct a set of unitaries that a worst-case learning algorithm can successfully distinguish but that only make minor differences when acting on states, so that a minimal number of queries have to be made in order to distinguish them. The upper bound is achieved by the average-case learning algorithm in Theorem 4 below when applied in the regime of exponentially small error.

Having established this no-go theorem for worst-case learning, we turn to a more realistic average-case learning alternative. Here, the accuracy is measured using the root-mean-square trace distance between output states over Haar-random inputs, $d_{\mathrm{avg}}(U, V) = \sqrt{\mathbb{E}_{|\psi\rangle}[d_{\mathrm{tr}}(U|\psi\rangle, V|\psi\rangle)^2]}$. This metric characterizes the average error when testing the learned unitary on randomly chosen input states.

We find that, similarly to the state-learning task, linear-in-G many queries are both necessary and sufficient to learn a unitary in the average case.

Theorem 4 (Average-case unitary learning). There exists an algorithm that learns an n-qubit unitary composed of G two-qubit gates to accuracy ϵ in the root-mean-square trace distance d_{avg} with high probability using $\tilde{\mathcal{O}}(G \min\{1/\epsilon^2, \sqrt{2^n}/\epsilon\})$ queries to the unknown unitary. Meanwhile, $\Omega(G/\epsilon)$ queries to the unitary or its inverse or the controlled versions are necessary for any such algorithm.

We show the upper bound in Appendix C 2 by combining a covering net with quantum hypothesis selection, similarly to the upper bound in Theorem 1. Our algorithm achieving the query complexity $\tilde{\mathcal{O}}(G/\epsilon^2)$ uses maximally entangled states and the Choi-Jamiołkowski duality [63–65]. With a bootstrap method similar to quantum phase estimation [23], we improve the ϵ dependence to the Heisenberg scaling $\tilde{\mathcal{O}}(1/\epsilon)$, albeit at the cost of a dimensional factor. It is an open question as to whether one can improve the ϵ dependence without incurring this dimensional factor. Without auxiliary systems, we prove a query-complexity bound of $\tilde{\mathcal{O}}(G \min\{1/\epsilon^4, (\sqrt{2^n})^3/\epsilon\})$. The lower bound is proven in Appendix C 3 by mapping to a fractional-query problem [23,66,67] and making use of a recent upper bound on the success probability in unitary distinguishing tasks [68]. In the case of learning generic unitaries, our result yields a $\Omega(4^n/\epsilon)$ lower bound, improving upon the $\Omega(4^n/n^2)$ bound from the recent work Ref. [56], which studies the hardness of learning Haar-(pseudo)random unitaries.

As Haar-random states are hard to generate in practice, we also discuss other input-state ensembles of physical interest. Relying on the equivalence of root-mean-square trace distances over different locally scrambled ensembles [69,70], recently established in Ref. [71], our algorithm achieves the same average-case guarantee over any such ensemble. Notable examples of locally scrambled ensembles include products of Haar-random single-qubit states or of random single-qubit stabilizer states, 2-designs on n-qubit states, and output states of random local quantum circuits with any fixed architecture.

The similar linear-in-G sample or query-complexity scaling in Theorems 1 and 4 hints at a common underlying source of complexity. However, in contrast to state learning, unitary learning comes with two natural such sources: (1) to readout input and output states and (2) to learn the mapping from inputs to outputs. The similarity between learning states and unitaries in terms of complexities suggests that the former may encapsulate the central difficulty in unitary learning, whereas the latter may be easy. This seemingly contradicts recent quantum no-free-lunch theorems [50,51,72], which state that $\Omega(2^n)$ samples are required to learn a generic unitary even from classical descriptions of input-output state pairs, highlighting the difficulty of (2).

To resolve this apparent contradiction, we reformulate the quantum no-free-lunch theorem (Theorem 17) from a unifying information-theoretic perspective in Appendix C 4. We highlight that enlarging the space for the classically described data allows us to systematically reduce the sample complexity until a single sample suffices to learn a general unitary. Therefore, the difficulty of learning the mapping, as indicated by quantum no-free-lunch theorems, vanishes when we allow auxiliary systems and query access to the unitary. Inspired by this observation, we give two ways of enlarging the representation space with auxiliary systems. The first is fundamentally quantum, making use of entangled input states [51]. The other is purely classical, relying on mixed-state inputs [73].

Theorem 5 (Learning with classical descriptions). There exists an algorithm that learns a generic n-qubit unitary with any nontrivial accuracy and with high success probability using $\mathcal{O}(2^n/r)$ classically described input-output pairs with mixed (entangled) input states of (Schmidt) rank r. Moreover, any such algorithm that is robust to noise needs at least $\Omega(2^n/r)$ samples.

Similarly to the case for state learning, our average-case unitary-learning algorithm is not computationally efficient. We show that this cannot be avoided. Under commonly believed cryptographic assumptions [43–48], any quantum algorithm that can learn unknown unitaries with circuit size $\tilde{\mathcal{O}}(G)$ to a small error in average-case distance from queries must have a computational time exponential in G. This implies the same computational hardness for worst-case unitary learning and a $\log n$ transition point of computational efficiency. Note that the hard instances that we construct are implementable with a similar number of Clifford and T gates [74]. Therefore, together with Theorem 2, this implies that there are no polynomial-time quantum algorithms for learning Clifford+T circuits with $\tilde{\omega}(\log n)$ T gates, answering an open question (the fifth question) in the survey in Ref. [49] negatively.

Theorem 6 (Unitary-learning computational complexity). Suppose that we are given N queries to an arbitrary unknown n-qubit unitary U consisting of $\tilde{\mathcal{O}}(G)$ two-qubit gates. Assume that RingLWE cannot be solved by a quantum computer in subexponential time. Then, any quantum algorithm that learns the unitary to within ϵ average-case distance d_{avg} must use $\exp(\Omega(\min\{G,n\}))$ time. Meanwhile, for $G=\mathcal{O}(\log n)$, the learning task can be solved in polynomial time.

C. Learning with physical functions

Apart from learning quantum states and dynamics themselves, a more classically minded learner may care more about learning classical functions resulting from quantum processes. We define these *physical functions* in Appendix D as functions $f(x,\{U_i\}_{i=1}^G,a)$ mapping $x\in[0,1]^\nu$ to $\mathbb{R}$ resulting from a physical experiment consisting of three steps: (1) a fixed-state-preparation procedure that can depend on x; (2) a unitary evolution consisting of G tunable two-qubit gates $\{U_i\}_{i=1}^G$ and arbitrary fixed unitaries that can depend on x, arranged in a circuit architecture a; and (3) the measurement of a fixed observable, the expectation of which is the function output. By tuning the local gates $\{U_i\}_{i=1}^G$ and potentially changing architecture a, we obtain a resulting class of functions that can be implemented in this general experimental setting. Despite the generality of this setup, we find that certain well-behaved functions are actually not physical in this sense: they cannot be efficiently approximated or learned via physical functions.

Theorem 7 (Approximating and learning with physical functions). To approximate and learn arbitrary 1-bounded and 1-Lipschitz $\mathbb{R}$-valued functions on $[0,1]^\nu$ to accuracy ϵ in $\|\cdot\|_\infty$ with high probability, using physical functions with G gates and variable circuit structures, we must use $G\geq\tilde{\Omega}(1/\epsilon^{\nu/2})$ gates and collect at least $\Omega(1/\epsilon^\nu)$ samples. If the circuit structure is fixed, we require $G\geq\tilde{\Omega}(1/\epsilon^\nu)$ gates.

We prove this in Appendix D by noting that to approximate arbitrary 1-bounded and 1-Lipschitz functions well, the complexity of experimentally implementable functions cannot be too small, as measured by the pseudodimension [75] or the fat-shattering dimension [76]. Then, the

gate-complexity lower bound follows because the function class complexity is limited by the circuit complexity [77] and we can appeal to results in classical learning theory [78] to obtain our sample-complexity lower bound.

It has been established that a classical neural network can learn to approximate any 1-bounded and 1-Lipschitz functions to accuracy ϵ in $\|\cdot\|_\infty$ with $\tilde{\Theta}(1/\epsilon^\nu)$ parameters, exponential in the number of variables ν, known as the curse of dimensionality [79]. Our results show that quantum neural networks can do no better. This result not only is relevant to the practical implementation of quantum machine learning, complementing existing results on the universal approximation of quantum neural networks [80–83], but also has deep implications for the physicality of the function class under consideration. It means that there are some many-variable 1-bounded and 1-Lipschitz functions that cannot be implemented in nature efficiently. On the other hand, certain more restricted function classes can be approximated using only $\mathcal{O}(1/\epsilon^2)$ parameters with both classical [79] and quantum neural networks [80], independent of the number of variables. This reveals a fundamental limitation on the functional expressivity of nature, practical QML models, and quantum signal processing algorithms [84,85]. We remark that while we prove this no-go result, achieving a quantum advantage may still be possible for other function classes [86].

III. NUMERICAL EXPERIMENTS

To support our theoretical findings, we conduct numerical experiments using our learning algorithm applied to pure states generated from G two-qubit gates. The results reflect the linear-in-G scaling of the sample complexity N from Theorem 1. We consider a large system size $n = 10\,000$ to illustrate that the sample complexity is independent of n. We study two families of unknown target states with different gate configurations: (a) the G gates are concentrated on 4 qubits; and (b) the G gates are randomly placed. We note that case (a) corresponds to the hard-to-learn states that we construct to prove the $\tilde{\Omega}(G)$ lower bound in Theorem 1, while in case (b) the gates are expected to spread out over the large system and form shallow circuits [87]. Due to the exponential-in-G computation time proved in Theorem 2, we restrict the gate size to $G = 10$ in case (a) and $G = 6$ in case (b). We perform the simulations by implementing the algorithm from Appendix B 1 using shallow Clifford classical shadows [88,89]. The details of the numerical implementation are provided in Appendix E.

The performance of our learning algorithm is shown in Fig. 2. Figure 2(a) corresponds to case (a) described in the previous paragraph and Fig. 2(b) corresponds to case (b). We provide contour plots of the average fidelity F of the reconstruction with different gate sizes G and sample sizes N. The sample complexity N with different gate sizes G is plotted in solid lines for different average fidelities F and in dashed lines for the median fidelity F_{med}. We see a linear dependence of sample complexity N on gate size G, in accordance with our theoretical bound $N = \tilde{\Theta}(G)$ from Theorem 1. Moreover, we note that a relatively small sample size $N \sim 50$ suffices to learn states with $G \sim 10$ gates on very large system size $n = 10\,000$, due to the fact that our sample complexity is independent of n.

IV. DISCUSSION

Our work provides a new and more fine-grained perspective on the fundamental problems of state and process tomography by analyzing them for the broad and physically relevant class of bounded-complexity states and

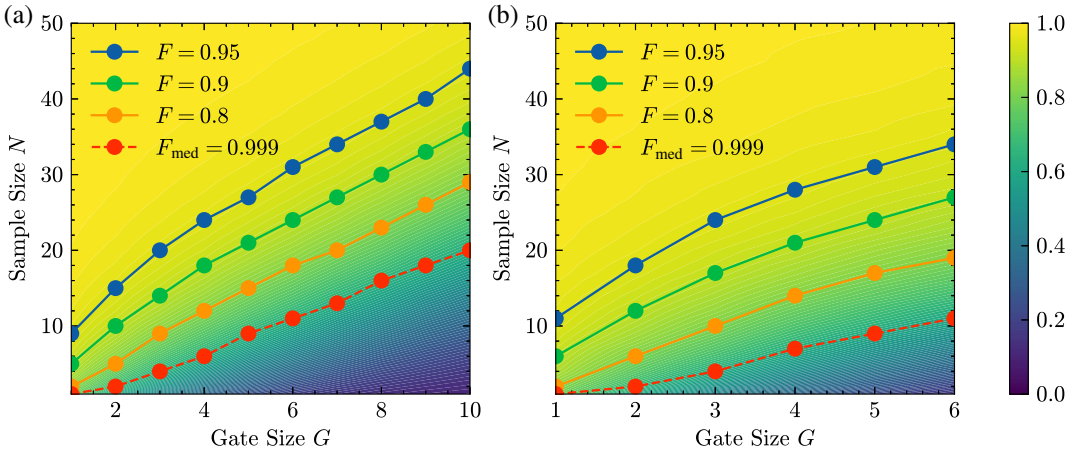

FIG. 2. The sample complexity N of the learning algorithm with different gate numbers G and reconstruction fidelities F. The unknown target states are pure states on $n = 10\,000$ qubits generated from G gates, either (a) concentrated on the first four qubits or (b) randomly placed. The contour plot represents the fidelity for different G and N averaged over many random instances. Sample complexities with average fidelity F and median fidelity F_{med} are plotted in solid and dashed lines, respectively.

unitaries. It complements existing literature on learning restricted classes of states or unitaries or their properties. Examples include stabilizer circuits and states [90–93], Clifford circuits with few non-Clifford T gates and their output states [74,94–97], matrix product operators [26] and states [98–100], phase states [91,101–103], permutationally invariant states [104–106], outputs of shallow quantum circuits [107], probably-approximately-correct (PAC) learning quantum states [108] and circuits [109], shadow tomography [42], classical shadow formalism [14,110–112], and property prediction of the outputs of quantum processes [113–115]. It also raises many interesting questions for future research.

First, to account for decoherence and imperfections in realistic experiments, it is natural to generalize our results to mixed states and channels. As our learning algorithms based on hypothesis selection and classical shadows rely on the purity or unitarity of the unknown state or process, it seems that different algorithmic approaches would be needed to go beyond states of constant rank. Moreover, while our results show that learners using only single-copy measurements and no coherent quantum processing can achieve optimal sample or query complexity (in G) for pure-state or unitary learning [in line with the state-tomography protocol in Ref. [22], which uses at most $\mathrm{rank}(\rho)$ copies at a time for the tomography of general state ρ], quantum enhanced learners, using multicopy measurements and coherent processing, may have an advantage in the case of mixed states and channels. Such a quantum advantage is known for general mixed-state tomography [116,117] and in certain channel-learning scenarios [113,115,118–122]; however, to our knowledge, not yet under assumptions of bounded complexity.

Second, there are several regimes of interest in which our results may be further extended. For instance, while we establish a computational efficiency transition for state and unitary learning at logarithmic circuit complexity, we leave open the question of computationally efficient learning with constraints beyond circuit complexity (e.g., constant-depth circuits where the gates are spread out). Another potential improvement related to the computational complexity is in regard to average-case computational hardness. While our computational lower bounds hold in the worst case, this does not tell us if most states or unitaries of bounded gate complexity are computationally hard to learn. Is there a worst-case to average-case reduction for this problem? Or, perhaps, is there an average-case notion of pseudorandomness that one could leverage here? An additional regime in which our work can be extended is as follows. Our adaptation of the bootstrap strategy from Ref. [23] to average-case unitary learning achieves Heisenberg scaling only at the cost of a dimension-dependent factor. Given recent work in state shadow tomography [123–125], it may not be possible to find a learner free from this dimensional factor while achieving the ϵ^{-1} scaling. Finding such a learner or disproving its existence could serve as an important contribution to recent progress on Heisenberg-limited learning in different scenarios [126–128].

Third, can we make learning even more efficient if the circuit structure is fixed and known in advance? Our upper bound already implies an algorithm with $\tilde{O}(G)$ sample complexity for fixed circuit structure but the lower-bound proof crucially relies on the ability to place gates freely in the construction of the packing net. A particular fixed-circuit structure of physical relevance is the brickwork circuit [129]. In Appendix F, we give preliminary results showing that if an n-qubit G-gate brickwork circuit suffices to implement an approximate unitary t-design [130], then the metric entropy of this unitary class with respect to d_{avg} is lower bounded by $\Omega(tn)$. Considering the known lower bound of $G \geq \tilde{\Omega}(tn)$ on the size of brickwork circuits implementing t-designs [130], the tightness of which is still an open problem [131], this may hint at a similar $\tilde{\Theta}(G)$ sample complexity of learning brickwork circuits.

Lastly, we outline a potential connection to the Brown-Susskind conjecture [132,133] originating from the wormhole-growth paradox in holographic duality [134–137]. Informally, the conjecture states that the complexity of a generic local quantum circuit grows linearly with the number of two-qubit gates for an exponentially long time, dual to the steady growth of the volume of a wormhole in the bulk theory. With "complexity" understood as "circuit complexity" [136], this conjecture has recently been confirmed for exact circuit complexity [138,139], while the case of approximate circuit complexity is only partially resolved [140,141]. Our work suggests an alternative approach to the Brown-Susskind conjecture. Namely, we have demonstrated that the complexity of learning quantum circuits grows linearly with the number of local gates in the worst case. If our bounds were extended to hold with high probability over random circuits with G gates, this would yield a sample-complexity version of the Brown-Susskind conjecture, suggesting the complexity of learning as a dual of the wormhole volume.

Via these open questions, tomography problems dating back to the early days of quantum computation and information connect closely to different avenues of current research in the field. Consequently, answering these questions will shed new light on fundamental quantum physics as well as on the frontiers of quantum complexity and quantum learning.

V. METHODS

In this section, we discuss the main ideas behind the proof of our results on the sample complexity of learning states (Theorem 1) and unitaries (Theorem 4), along with the computational complexity (Theorems 2 and 6).

A. Sample-complexity upper bounds

We prove the upper bounds in Theorems 1 and 4, using a hypothesis-selection protocol similar to that of Ref. [54] but now based on classical shadow tomography [110] that enables a linear-in-G scaling.

1. State learning

For state learning, we first take a minimal covering net $\mathcal{N}$ over the set of states with bounded circuit complexity G such that for any such state $|\psi\rangle$, there exists a state in the covering net that is ϵ-close to $|\psi\rangle$ in trace distance. This net then serves as a set of candidate states, from which the learning algorithm will select one. Importantly, we prove that the cardinality of $\mathcal{N}$ can be upper bounded by $|\mathcal{N}| \leq e^{\tilde{\mathcal{O}}(G)}$. Here, note that the tilde hides a logarithmic factor in terms of system size, which we remove using a more detailed analysis with ideas from junta learning [55].

Next, we use classical shadows created via random Clifford measurements [110] to estimate the trace distance between the unknown state and each of the candidates in $\mathcal{N}$. This is achieved by estimating the expectation value of the Helstrom measurement [142], which is closely related to the trace distance between two states. As the rank of Helstrom measurements between pure states is at most 2, Clifford classical shadows can efficiently estimate all $\binom{|\mathcal{N}|}{2}$ of them simultaneously to ϵ error using $\mathcal{O}(\log|\mathcal{N}|/\epsilon^2) \leq \tilde{\mathcal{O}}(G/\epsilon^2)$ copies of $|\psi\rangle$. Then, we select the candidate that has the smallest trace distance from $|\psi\rangle$ as the output.

The above strategy leads to a sample-complexity upper bound that depends logarithmically on the number of qubits n. This is undesirable when the circuit complexity G is smaller than $n/2$ (i.e., when some of the qubits are in fact never influenced by the circuit). We improve our algorithm in this small-size regime by first performing a junta learning step [55] to identify which of the qubits are acted on nontrivially. After that, we enhance our protocol with a measure-and-postselect step. This allows us to construct a covering net only over the qubits acted upon nontrivially, the cardinality of which no longer depends on n. We then perform the hypothesis selection as before. In this way, we are able to achieve a sample complexity independent of system size.

2. Unitary learning

The algorithm for unitary learning is similar to the state-learning protocol. When allowing the use of an auxiliary system, we utilize the fact that the average-case distance between unitaries is equivalent to the trace distance between their Choi states. In this way, we can reduce the problem to state learning of the Choi states and achieve the $\tilde{\mathcal{O}}(G/\epsilon^2)$ sample complexity. Without auxiliary systems, we can sample random input states and perform one-shot Clifford shadows on the outputs to estimate the squared average-case distance, resulting in an $\tilde{\mathcal{O}}(G/\epsilon^4)$ sample complexity with a suboptimal ϵ dependence.

Furthermore, we improve the ϵ dependence in unitary learning to the Heisenberg scaling $\tilde{\mathcal{O}}(1/\epsilon)$ via a bootstrap method similar to that of Ref. [23], using the above learning algorithm as a subroutine. Specifically, we iteratively refine our learning outcome $\hat{U}$ by performing hypothesis selection over a covering net of $(U\hat{U}^\dagger)^p$, with p increasing exponentially as the iteration proceeds. Although the circuit complexity of $(U\hat{U}^\dagger)^p$ grows with p, a covering net with p-independent cardinality can be constructed based on the one-to-one correspondence to U. However, unlike the diamond-distance learner considered in Ref. [23], which has fine control over every eigenvalue of the unitaries, our average-case learner only has control over the average of the eigenvalues. Thus for the bootstrap to work (i.e., for the learning error to decrease with increasing p), the average-case learner has to work in an exponentially small error regime, which results in a dimensional factor in the final sample complexity $\tilde{\mathcal{O}}(\sqrt{2^n}G/\epsilon)$.

B. Sample-complexity lower bounds

We prove the sample-complexity lower bounds in Theorems 1 and 4 by reduction to distinguishing tasks. Specifically, if we can learn the state or unitary to within ϵ error, then we can use this learning algorithm to distinguish a set of states or unitaries that are 3ϵ far apart from each other. Hence a lower bound on the sample complexity of distinguishing states or unitaries from a packing net implies a lower bound for the learning task.

1. State learning

For state learning, we construct a packing net $\mathcal{M}$ of the set of $(\log_2 G)$-qubit states, which we later tensor product with zero states on the remaining qubits. These states have circuit complexity of approximately G because $\mathcal{O}(2^k)$ two-qubit gates can implement any pure k-qubit states [143]. We prove that the cardinality of $\mathcal{M}$ can be lower bounded by $e^{\Omega(G)}$. This means that to distinguish the states in $\mathcal{M}$, one has to gather $\Omega(\log|\mathcal{M}|) \geq \Omega(G)$ bits of information. Meanwhile, Holevo's theorem [144] asserts that the amount of information carried by each sample is upper bounded by $\tilde{\mathcal{O}}(\epsilon^2)$ [145]. Hence, we need at least $\tilde{\Omega}(G/\epsilon^2)$ copies of the unknown state.

2. Unitary learning

Similarly, for unitary learning, we construct a packing net by stacking all the gates into $\log_4 G$ qubits, using the fact that $\mathcal{O}(4^k)$ two-qubit gates suffice to implement any k-qubit unitaries [146]. Lacking an analogue of Holevo's theorem for unitary queries, we turn to a recently established bound on the success probability of unitary discrimination [68] and obtain an $\Omega(G)$ sample-complexity lower

bound for constant ϵ. To incorporate the ϵ dependence, we follow Ref. [23] and map the problem into a fractional-query problem. We show that with N queries, we can use the learning algorithm to simulate [66,67] an $\mathcal{O}(\epsilon N)$ query algorithm that solves the above constant-accuracy distinguishing problem. This gives us the desired $N \geq \Omega(G/\epsilon)$ lower bound.

C. Computational hardness

We prove the computational-complexity lower bounds in Theorems 2 and 6, again by reduction to distinguishing tasks, the hardness of which relies on cryptographic primitives in this case. In particular, we show that if we can learn the state or unitary in polynomial time, then we can use this learning algorithm to efficiently distinguish between pseudorandom states and functions [147,148] and truly random states and functions. We note that similar ideas have been used to establish a cryptographic no-cloning theorem [147] for pseudorandom quantum states (PRSs) but without gate-complexity dependence and the unitary counterpart. The RingLWE hardness assumption here may also be relaxed to the existence of appropriate quantum secure PRS-PRF constructions that have the same gate complexity, as discussed below.

Our proofs rely on the construction of quantum secure pseudorandom functions (PRFs) that can be implemented using TC^0 circuits, subject to the assumption that RingLWE cannot be solved by a quantum computer in subexponential time [45]. We show that the circuit construction of Ref. [45] can be implemented quantumly using $G = \mathcal{O}(n\mathrm{polylog}(n))$ gates by converting this TC^0 circuit into a quantum circuit that computes the same function. With this construction, we can prove the computational hardness of learning when $G = \mathcal{O}(n\mathrm{polylog}(n))$ as follows.

1. State learning

For state learning, we utilize these quantum secure PRFs to construct PRSs—in particular, binary-phase states from Refs. [147,149]—with $G = \mathcal{O}(n\mathrm{polylog}(n))$ gates. Given copies of some unknown quantum state that is promised to either be a PRS or a Haar-random state, we design a procedure that can distinguish between these two cases. The distinguisher uses our algorithm for learning states along with the SWAP test applied to the learned state and the given state [150,151]. Thus, we show that if our learning algorithm was able to computationally efficiently learn PRSs, then we would have an efficient distinguisher between PRS and Haar-random states, contradicting the definition of a PRS [147].

2. Unitary learning

The proof idea in the unitary setting is similar. In this case, we consider PRFs directly rather than the PRS construction. Given query access to some unknown unitary that is promised to be the unitary oracle of either a PRF or a uniformly random Boolean function, we design a procedure that can distinguish between these two cases. The distinguisher uses our algorithm for learning unitaries along with the SWAP test [150,151]. Here, we query the given or learned unitaries on a random tensor product of single-qubit stabilizer states and conduct the SWAP test between the output states. In this way, we show that if our learning algorithm was able to computationally efficiently learn a unitary implementing a PRF, then we would have an efficient distinguisher between PRFs and uniformly random functions, which contradicts the definition of a PRF [148].

We then go one step further and show computational hardness for circuit size $\tilde{\mathcal{O}}(G)$. To do this, we rely critically on the assumption that RingLWE is hard not just to polynomial-time quantum algorithms but even to quantum algorithms that run for longer (subexponential) time. This allows us to take a much smaller input size to the PRS or PRF in our previous constructions (i.e., over $\mathcal{O}(G)$ qubits, which can be implemented with $\tilde{\mathcal{O}}(G)$ gates). The subexponential computational hardness of RingLWE then implies that solving the learning tasks requires time exponential in G.

Meanwhile, for $G = \mathcal{O}(\log n)$, the learning tasks can be solved efficiently by junta learning and standard tomography methods. This establishes $\log n$ circuit complexity as a transition point of computational efficiency. This also implies that the circuit complexity of the PRS-PRF constructions in Refs. [45,149] is optimal up to logarithmic factors; otherwise, it would contradict the efficient tomography of $\mathcal{O}(\log n)$-complexity states or unitaries. Finally, we note that the PRS or PRF that we consider can be implemented with a similar number of Clifford and T gates, extending our results to Clifford+T circuits.

The code that generates the data presented in the figures and that supports the other findings of this study is available at Ref. [152].

ACKNOWLEDGMENTS

We thank John Preskill for valuable feedback and continuous support throughout this project. We also thank Ryan Babbush, Yiyi Cai, Charles Cao, Alexandru Gheorghiu, András Gilyén, Alex B. Grilo, Jerry Huang, Minghao Liu, Nadine Meister, Chris Pattison, Alexander Poremba, Xiao-Liang Qi, Ruohan Shen, Mehdi Soleimanifar, Yu Tong, Tzu-Chieh Wei, Tai-Hsuan Yang, and Nengkun Yu for fruitful discussions. Finally, we thank Antonio Anna Mele and the anonymous reviewers for feedback on an earlier version of this paper. H.Z. was supported by a Caltech Summer Undergraduate Research Fellowship (SURF). L.L. was supported by a Mellon Mays Undergraduate Fellowship. H.H. was supported by a Google Ph.D. Fellowship and a MediaTek Research Young

Scholarship. H.H. acknowledges the visiting associate position at the Massachusetts Institute of Technology. This work is supported by a collaboration between the U.S. Department of Energy (DOE) and other agencies. Y.Q. acknowledges financial support from the Quantum Systems Accelerator program funded by the National Quantum Information Science Research Centers under the DOE Office of Science. M.C.C. was supported by a Deutscher Akademischer Austauschdienst (DAAD) Postdoctoral Researchers International Mobility Experience (PRIME) fellowship. This work was done (in part) while a subset of the authors were visiting the Simons Institute for the Theory of Computing. The Institute for Quantum Information and Matter is a National Science Foundation Physics Frontiers Center.

Y.Q., H.H., and M.C.C. conceived the project. H.Z., L.L., and I.K. led the development of the theory and the analytic calculations. H.Z., I.K., and H.H. designed and conducted the numerical experiments. All authors contributed to the mathematical aspects of this work. H.Z. and L.L. wrote the manuscript, with input from all authors.

The authors declare no competing interests.

APPENDIX A: PRELIMINARIES

Throughout the appendixes, we use $d = 2^n$ to denote the dimension of the n-qubit Hilbert space unless otherwise stated.

1. Distance metrics

Here, we review some distance metrics and their properties used throughout our proofs. In the main text, we have already introduced the trace distance

$$d_{\mathrm{tr}}(|\psi\rangle, |\phi\rangle) = \frac{1}{2}\| |\psi\rangle\langle\psi| - |\phi\rangle\langle\phi| \|_1, \tag{A1}$$

which is analogously defined for density matrices as $d_{\mathrm{tr}}(\rho,\sigma) = \|\rho-\sigma\|_1/2$, the diamond distance

$$d_\diamond(U,V) = \max_\rho \|(U\otimes I)\rho(U\otimes I)^\dagger - (V\otimes I)\rho(V\otimes I)^\dagger\|_1, \tag{A2}$$

and the root-mean-square trace distance

$$d_{\mathrm{avg}}(U,V) = \sqrt{\mathbb{E}_{|\psi\rangle}[d_{\mathrm{tr}}(U|\psi\rangle, V|\psi\rangle)^2]} \tag{A3}$$

where the expectation is taken over the Haar measure [153].

Apart from these, we also use the following auxiliary distance metrics. We define the quotient spectral distance

$$d_2'(U,V) = \min_{e^{i\phi}\in U(1)} \|U - e^{i\phi}V\| \tag{A4}$$

to be the spectral distance $d_2(U,V) = \|U-V\|$ up to a global phase. Similarly, we define the quotient-normalized Frobenius distance

$$d_F'(U,V) = \min_{e^{i\phi}\in U(1)} \frac{1}{\sqrt{d}}\|U - e^{i\phi}V\|_F \tag{A5}$$

as the normalized Frobenius-norm distance $d_F(U,V) = \frac{1}{\sqrt{d}}\|U-V\|_F$ up to a global phase.

The following lemma shows that the (quotient) spectral distance and the diamond distance are equivalent.

Lemma 1 (Spectral and diamond distance of unitaries; variant of Ref. [154, Lemma B.5]). For any two d-dimensional unitaries U and V, we have

$$\frac{1}{\sqrt{2}}d_2'(U,V) \le \frac{1}{2}d_\diamond(U,V) \le d_2'(U,V) \le \|U-V\|. \tag{A6}$$

Proof. Since stabilization is not necessary for computing the diamond distance of two unitary channels [155], we have

$$\begin{aligned}
&\frac{1}{2}d_\diamond(U,V)\\
&= \max_{|\psi\rangle}\frac{1}{2}\|U|\psi\rangle\langle\psi|U^\dagger - V|\psi\rangle\langle\psi|V^\dagger\|_1\\
&= \max_{|\psi\rangle}\sqrt{1-|\langle\psi|U^\dagger V|\psi\rangle|^2}\\
&= \max_{|\psi\rangle}\sqrt{(1+|\langle\psi|U^\dagger V|\psi\rangle|)(1-|\langle\psi|U^\dagger V|\psi\rangle|)}\\
&\ge \max_{|\psi\rangle}\frac{1}{\sqrt{2}}\sqrt{2(1-|\langle\psi|U^\dagger V|\psi\rangle|)}\\
&= \frac{1}{\sqrt{2}}\min_{e^{i\phi}\in U(1)}\max_{|\psi\rangle}\|U|\psi\rangle - e^{i\phi}V|\psi\rangle\|_2\\
&= \frac{1}{\sqrt{2}}\min_{e^{i\phi}\in U(1)}\|U - e^{i\phi}V\| = \frac{1}{\sqrt{2}}d_2'(U,V),
\end{aligned} \tag{A7}$$

where we have used $|\langle\psi|U^\dagger V|\psi\rangle| \ge 0$ and the standard conversion between the trace distance and the fidelity. This proves the first inequality. Similarly, we have

$$\begin{aligned}
&\frac{1}{2}d_\diamond(U,V)\\
&= \max_{|\psi\rangle}\frac{1}{2}\|U|\psi\rangle\langle\psi|U^\dagger - V|\psi\rangle\langle\psi|V^\dagger\|_1\\
&= \max_{|\psi\rangle}\sqrt{1-|\langle\psi|U^\dagger V|\psi\rangle|^2}\\
&= \max_{|\psi\rangle}\sqrt{(1+|\langle\psi|U^\dagger V|\psi\rangle|)(1-|\langle\psi|U^\dagger V|\psi\rangle|)}\\
&\le \max_{|\psi\rangle}\sqrt{2(1-|\langle\psi|U^\dagger V|\psi\rangle|)}
\end{aligned}$$

$$= \min_{e^{i\phi}\in U(1)} \max_{|\psi\rangle} \|U|\psi\rangle - e^{i\phi}V|\psi\rangle\|_2$$
$$= \min_{e^{i\phi}\in U(1)} \|U - e^{i\phi}V\| = d_2'(U,V), \tag{A8}$$

where we have used $\left|\langle\psi|U^\dagger V|\psi\rangle\right| \le 1$, proving the second inequality. The third inequality follows immediately from $d_2'(U,V) = \min_{e^{i\phi}\in U(1)} \|U - e^{i\phi}V\| \le \|U - V\|$. ■

We will also utilize the subadditivity of the diamond distance.

Lemma 2 (Subadditivity of diamond distance [156, Prop. 3.48]). For any d-dimensional unitaries U_1, U_2, V_1, V_2, we have the following inequality:

$$d_\diamond(U_2U_1, V_2V_1) \le d_\diamond(U_2, V_2) + d_\diamond(U_1, V_1). \tag{A9}$$

From the standard relationship between different p-norms, we have the following relation between d_2' and d_F'.

Lemma 3 (Norm conversion between quotient spectral and normalized Frobenius distance). For any two d-dimensional unitaries U and V, we have

$$\frac{1}{\sqrt{d}} d_2'(U,V) \le d_F'(U,V) \le d_2'(U,V). \tag{A10}$$

Proof. For any $e^{i\phi} \in U(1)$, the standard relation between matrix norms gives us

$$\|U - Ve^{i\phi}\| \le \|U - Ve^{i\phi}\|_F \le \sqrt{d}\|U - Ve^{i\phi}\|. \tag{A11}$$

Taking the minimum of $\|U - Ve^{i\phi}\|_F$ over $e^{i\phi}$ in the first inequality and dividing by $\sqrt{d}$, we obtain

$$\frac{1}{\sqrt{d}} d_2'(U,V) \le \frac{1}{\sqrt{d}}\|U - Ve^{i\phi}\| \le d_F'(U,V). \tag{A12}$$

Similarly, taking the minimum of $\|U - Ve^{i\phi}\|$ over $e^{i\phi}$ in the second inequality and dividing by $\sqrt{d}$ yields

$$d_F'(U,V) \le \frac{1}{\sqrt{d}}\|U - Ve^{i\phi}\|_F \le d_2'(U,V). \tag{A13}$$

Thus we have the desired results. ■

The following lemma collects some useful properties of d_F' and, in particular, shows that d_F' and d_{avg} are equivalent.

Lemma 4 (Properties of quotient-normalized Frobenius distance). For any two d-dimensional unitaries U and V, we have the following:

(1) $\frac{1}{2} d_F'(U,V) \le d_{\text{avg}}(U,V) \le d_F'(U,V)$.

(2) For any integer $p \ge 1$, $d_F'(U^p, V^p) \le p d_F'(U,V)$.

(3) For any integer $p \ge 1$, if $d_F'(U,I), d_F'(V,I) \le (4/(25\pi)/\sqrt{d})$, then $d_F'(U^{1/p}, V^{1/p}) \le (2/p)\, d_F'(U,V)$.

Item (3) can be viewed as a version of Ref. [23, Lemma 3.1].

Proof. (1) From properties of the Haar integral (see, e.g., Ref. [157, Example 50]), we have

$$d_{\text{avg}}(U,V)^2 = 1 - \frac{d + |\text{tr}(U^\dagger V)|^2}{d(d+1)}. \tag{A14}$$

On the other hand, we have

$$d_F'^2(U,V) = \min_{e^{i\phi}\in U(1)} \frac{1}{d}\|U - Ve^{i\phi}\|_F^2$$
$$= \min_{e^{i\phi}\in U(1)} 2 - \frac{2}{d}\text{Re}[\text{tr}(U^\dagger V e^{i\phi})]$$
$$= 2 - \frac{2}{d}|\text{tr}(U^\dagger V)|. \tag{A15}$$

Combining them, we obtain

$$d_{\text{avg}}(U,V)^2 = \frac{d}{d+1} d_F'^2(U,V)\left(1 - \frac{d_F'^2(U,V)}{4}\right)$$
$$\in \left[\frac{1}{4} d_F'^2(U,V), d_F'^2(U,V)\right], \tag{A16}$$

because $d_F'^2(U,V) \in [0,2]$. Thus we have established item (1).

(2) From the triangle inequality, we have

$$d_F'(U^p, V^p) \le \sum_{k=1}^{p} d_F'(U^{p+1-k}V^{k-1}, U^{p-k}V^k)$$
$$= \sum_{k=1}^{p} d_F'(U,V) = p d_F'(U,V), \tag{A17}$$

where we have used the unitary invariance of d_F'. This proves item (2).

(3) We first prove the following modified version without the global phase: "If $d_F(U,I), d_F(V,I) \le (4/(5\pi)/\sqrt{d})$, then $d_F(U^{1/p}, V^{1/p}) \le (2/p) d_F(U,V)$." Let $U = e^X, V = e^Y$ with $\|X\|, \|Y\| \le \pi$. We can refine the bound on $\|X\|, \|Y\|$ by noting the following:

$$\|X\| \le \frac{\pi}{2}\|e^X - I\| \le \frac{\pi}{2}\|U - I\|_F$$
$$= \frac{\pi\sqrt{d}}{2} d_F(U,I) \le \frac{2}{5}, \tag{A18}$$

where the first inequality can be seen from eigenvalue analysis as follows. Let $i\theta_k$ be the eigenvalues of X with

$|\theta_k| \le \pi$. Then, we have

$$\|X\| = \max_k |\theta_k| \le \pi \max_k \left|\sin\frac{\theta_k}{2}\right| = \frac{\pi}{2}\max_k \left|e^{i\theta_k} - 1\right| = \frac{\pi}{2}\|e^X - I\|. \tag{A19}$$

Similarly, we have $\|Y\| \le 2/5$.

Next, we prove the following inequality when $\|X\|, \|Y\| \le 2/5$ (similar to Ref. [158, Appendix D]):

$$\frac{1}{2}\|X - Y\|_F \le \|e^X - e^Y\|_F \le \|X - Y\|_F. \tag{A20}$$

For the upper bound, we use the triangle inequality and a telescoping sum representation. For any $m \in \mathbb{N}$,

$$\|e^X - e^Y\|_F \le \sum_{k=1}^{m} \|e^{(k-1)X/m}(e^{X/m} - e^{Y/m})e^{(m-k)Y/m}\|_F = m\|e^{X/m} - e^{Y/m}\|_F, \tag{A21}$$

and by taking $m \to \infty$ we arrive at the upper bound. For the lower bound, note that by the triangle inequality, we have

$$\|e^X - e^Y\|_F = \left\|\sum_{k=1}^{\infty}\frac{1}{k!}(X^k - Y^k)\right\|_F \ge \|X - Y\|_F - \left\|\sum_{k=2}^{\infty}\frac{1}{k!}(X^k - Y^k)\right\|_F. \tag{A22}$$

The second term can be upper bounded by

$$\left\|\sum_{k=2}^{\infty}\frac{1}{k!}(X^k - Y^k)\right\|_F = \left\|\sum_{k=2}^{\infty}\sum_{l=1}^{k}\frac{1}{k!}X^{l-1}(X - Y)Y^{k-l}\right\|_F \le \sum_{k=2}^{\infty}\frac{k}{k!}\left(\frac{2}{5}\right)^k\|X - Y\|_F = (e^{2/5} - 1)\|X - Y\|_F, \tag{A23}$$

where we have used $\|AB\|_F \le \|A\| \cdot \|B\|_F$ and $\|X\|, \|Y\| \le 2/5$. Plugging this bound back in, we arrive at the lower bound

$$\|e^X - e^Y\|_F \ge (2 - e^{2/5})\|X - Y\|_F \ge \frac{1}{2}\|X - Y\|_F. \tag{A24}$$

Equation (A20) in particular implies

$$d_F(U^{1/p}, V^{1/p}) \le \frac{1}{p\sqrt{d}}\|X - Y\|_F \le \frac{2}{p}d_F(U, V), \tag{A25}$$

and thus the modified version of our claim.

Finally, we deal with the global phase and prove the d'_F version, where we assume that $d'_F(U, I), d'_F(V, I) \le (4/(25\pi)/\sqrt{d})$. Let $e^{i\phi_U}, e^{i\phi_V}, e^{i\phi} \in U(1)$ denote the global phases that minimize $d_F(U, Ie^{i\phi_U}), d_F(V, Ie^{i\phi_V})$ and $d_F(Ue^{-i\phi_U}, Ve^{-i\phi_V}e^{i\phi})$, respectively. Then, $d_F(U, Ie^{i\phi_U})$, $d_F(V, Ie^{i\phi_V}) \le (4/(25\pi)/\sqrt{d})$ by assumption, and $d_F(Ue^{-i\phi_U}, Ve^{-i\phi_V}) \le d_F(U, Ie^{i\phi_U}) + d_F(V, Ie^{i\phi_V}) \le (8/(25\pi)/\sqrt{d})$. Therefore,

$$\begin{aligned} d_F(e^{i\phi}, I) &\le d_F(e^{i\phi}, (Ve^{-i\phi_V})^\dagger(Ue^{-i\phi_U})) \\ &\quad + d_F((Ve^{-i\phi_V})^\dagger(Ue^{-i\phi_U}), I) \\ &= d_F(Ue^{-i\phi_U}, Ve^{-i\phi_V}e^{i\phi}) + d_F(Ue^{-i\phi_U}, Ve^{-i\phi_V}) \\ &\le 2d_F(Ue^{-i\phi_U}, Ve^{-i\phi_V}) \\ &\le \frac{16/(25\pi)}{\sqrt{d}}. \end{aligned} \tag{A26}$$

This means that $d_F(Ue^{-i\phi_U}e^{-i\phi}, I) \le d_F(U, Ie^{i\phi_U}) + d_F(e^{i\phi}, I) \le ((4+16)/(25\pi)/\sqrt{d}) = (4/(5\pi)/\sqrt{d})$. We also know that $d_F(Ve^{-i\phi_V}, I) \le (4/(25\pi)/\sqrt{d}) \le (4/(5\pi)/\sqrt{d})$. Thus the two matrices $Ue^{-i\phi_U}e^{-i\phi}$ and $Ve^{-i\phi_V}$ satisfy the condition of the modified version without global phase and we thus have

$$\begin{aligned} d'_F(U^{1/p}, V^{1/p}) &\le d'_F(U^{1/p}, V^{1/p}(e^{-i\phi_V})^{1/p}(e^{i\phi_U})^{1/p}(e^{i\phi})^{1/p}) \\ &= d_F((Ue^{-i\phi_U}e^{-i\phi})^{1/p}, (Ve^{-i\phi_V})^{1/p}) \\ &\le \frac{2}{p}d_F(Ue^{-i\phi_U}e^{-i\phi}, Ve^{-i\phi_V}) = d'_F(U, V). \end{aligned} \tag{A27}$$

This concludes the proof of item (3). ■

Haar-random states are in general hard to generate. One may want to use other ensembles of input states and the associated distance metric for average-case learning. A class of ensembles of physical interest is that of locally scrambled ensembles [69,70], defined as follows.

Definition 1 (Locally scrambled ensembles up to the second moment). An ensemble $\mathcal{S}$ of (i.e., a distribution over) n-qubit states is called a locally scrambled ensemble up to the second moment if it is of the form $\mathcal{S} = \mathcal{U}|0\rangle^{\otimes n}$, where $\mathcal{U}$ is an ensemble of unitaries that is locally scrambled up to the second moment. That is, there exists another unitary ensemble $\mathcal{U}'$, such that: (1) for any U' randomly sampled from $\mathcal{U}'$ and for any tensor product of single-qubit unitaries $\otimes_{i=1}^n U_i$, $U'\otimes_{i=1}^n U_i$ follows the same distribution of $\mathcal{U}'$; and (2) for any $2n$-qubit density matrices ρ, we have $\mathbb{E}_{U\sim\mathcal{U}}[U^{\otimes 2}\rho(U^\dagger)^{\otimes 2}] = \mathbb{E}_{U'\sim\mathcal{U}'}[U'^{\otimes 2}\rho(U'^\dagger)^{\otimes 2}]$. We use $\mathbb{S}_{\mathrm{LS}}^{(2)}$ to denote the set of all such state ensembles.

Notable examples of these ensembles include n-qubit Haar-random states, products of Haar-random single-qubit

states, products of random single-qubit stabilizer states, 2-designs on n-qubit states, and output states of random local quantum circuits with any fixed architecture. The following lemma from the study of out-of-distribution generalization [71] shows that these ensembles lead to mutually equivalent average-case distance metrics.

Lemma 5 (Equivalence of locally scrambled average-case distances [71, Theorem 1]). We denote by $d_P(U,V) = \sqrt{\mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(U|\psi\rangle, V|\psi\rangle)^2]}$ the root-mean-square trace distance with respect to an ensemble P. For any $P, Q \in \mathbb{S}_{\mathrm{LS}}^{(2)}$ and for any unitaries U, V, we have

$$\frac{1}{\sqrt{2}} d_Q(U,V) \leq d_P(U,V) \leq \sqrt{2} d_Q(U,V). \tag{A28}$$

The following lemma shows that the triangle inequality holds for d_P (and, in particular, d_{avg}).

Lemma 6 (Triangle inequality for average-case distance). Let $d_P(U,V) = \sqrt{\mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(U|\psi\rangle, V|\psi\rangle)^2]}$ be the root-mean-square trace distance with respect to an ensemble P. For any three unitaries U, V, and W, we have the triangle inequality

$$d_P(U,V) \leq d_P(U,W) + d_P(W,V). \tag{A29}$$

Proof. Note that

$$\begin{aligned} d_P^2(U,V) &= \mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(U|\psi\rangle, V|\psi\rangle)^2] \leq \mathbb{E}_{|\psi\rangle\sim P}[(d_{\mathrm{tr}}(U|\psi\rangle, W|\psi\rangle) + d_{\mathrm{tr}}(W|\psi\rangle, V|\psi\rangle))^2] \\ &= d_P^2(U,W) + d_P^2(W,V) + 2\mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(U|\psi\rangle, W|\psi\rangle) d_{\mathrm{tr}}(W|\psi\rangle, V|\psi\rangle)] \\ &\leq d_P^2(U,W) + d_P^2(W,V) + 2\sqrt{\mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(U|\psi\rangle, W|\psi\rangle)^2]} \cdot \sqrt{\mathbb{E}_{|\psi\rangle\sim P}[d_{\mathrm{tr}}(W|\psi\rangle, V|\psi\rangle)^2]} \\ &= (d_P(U,W) + d_P(W,V))^2, \end{aligned} \tag{A30}$$

where we have used the triangle inequality for d_{tr} and the Cauchy-Schwartz inequality. Taking the square root gives us the desired result. ■

2. Covering and packing nets

Our results in state and unitary learning utilize a tool from high-dimensional probability theory, namely, covering and packing nets. We employ covering nets in our proofs of the sample-complexity upper bounds and packing nets in our proofs of sample-complexity lower bounds. Intuitively, covering and packing nets characterize the complexity of a space by discretizing it with small balls of a given resolution. We formally define these concepts below.

Definition 2 (Covering net or number and metric entropy). Let (X,d) be a metric space. Let $K \subseteq X$ be a subset and $\epsilon > 0$. Then, define the following:

(1) $N \subseteq K$ is an *ϵ-covering net* of K if for any $x \in K$, there exists a $y \in N$ such that $d(x,y) \leq \epsilon$.
(2) The *covering number* $\mathcal{N}(K,d,\epsilon)$ of K is the smallest possible cardinality of an ϵ-covering net of K.
(3) The *metric entropy* is $\log \mathcal{N}(K,d,\epsilon)$.

We can similarly define a packing net.

Definition 3 (Packing net or number). Let (X,d) be a metric space. Let $K \subseteq X$ be a subset and $\epsilon > 0$. Then, define the following:

(1) $N \subseteq K$ is an *ϵ-packing net* of K if for any $x, y \in N$, $d(x,y) > \epsilon$.
(2) The *packing number* $\mathcal{M}(K,d,\epsilon)$ of K is the largest possible cardinality of an ϵ-packing net of K.

The following equivalence between covering and packing numbers is often useful.

Lemma 7 (Covering and packing are equivalent [53, Section 4.2]). Let (X,d) be a metric space. Let $K \subseteq X$ and $\epsilon > 0$. We have

$$\mathcal{N}(K,d,\epsilon/2) \geq \mathcal{M}(K,d,\epsilon) \geq \mathcal{N}(K,d,\epsilon). \tag{A31}$$

Covering numbers also have the following monotonicity property.

Lemma 8 (Monotonicity of covering number [53, Section 4.2]). Let (K,d) be a metric space. If $L \subseteq K$, then $\mathcal{N}(L,d,\epsilon) \leq \mathcal{N}(K,d,\epsilon/2)$.

For our purposes, we need the following upper and lower bounds on the covering number of the unitary group.

Since the states that we consider can be generated by unitaries applied to a fixed input state, a covering-number upper bound for unitaries with respect to the diamond distance implies a corresponding covering-number upper bound for states with respect to the trace distance.

Lemma 9 (Covering number of the unitary group (Refs. [159, Proposition 7). , [158, Lemma 1], and [154, Lemma C.1])] Let $\|\cdot\|'$ be any unitarily invariant norm. There exist universal constants $c_1, c_2 > 0$ such that for any $\epsilon \in (0, 2]$, the covering number of the d-dimensional unitary group $U(d)$ with respect to the norm $\|\cdot\|'$ satisfies

$$\left(\frac{c_1}{\epsilon}\right)^{d^2} \leq \mathcal{N}(U(d), \|\cdot\|', \|I\|'\epsilon) \leq \left(\frac{c_2}{\epsilon}\right)^{d^2}. \tag{A32}$$

In particular, for the spectral norm $\|\cdot\|$, we have the upper bound $\mathcal{N}(U(d), \|\cdot\|, \epsilon) \leq (6/\epsilon)^{2d^2}$. For the Frobenius norm $\|\cdot\|_F$, we have $(c_1/\epsilon)^{d^2} \leq \mathcal{N}(U(d), \|\cdot\|_F, \sqrt{d}\epsilon) \leq (c_2/\epsilon)^{d^2}$.

We can use this result to bound the covering number for n-qubit unitaries consisting of G two-qubit gates.

Theorem 8 (Covering number of G-gate unitaries). Let $U^G \subseteq U(2^n)$ be the set of n-qubit unitaries that can be implemented by G two-qubit gates. Then, for any $\epsilon \in (0, 1]$, there exist universal and positive constants c_1, c_2, and C such that for $1 \leq G/C \leq 4^{n+1}$, the metric entropy of U^G with respect to the normalized Frobenius distance d_F can be bounded as

$$\begin{aligned}\frac{G}{4C}\log\left(\frac{c_1}{\epsilon}\right) &\leq \log\mathcal{N}(U^G, d_F, \epsilon) \\ &\leq 16G\log\left(\frac{c_2 G}{\epsilon}\right) + 2G\log n.\end{aligned} \tag{A33}$$

Moreover, the metric entropy with respect to the diamond distance $d_\diamond$ can be explicitly upper bounded by

$$\log\mathcal{N}(U^G, d_\diamond, \epsilon) \leq 32G\log\left(\frac{12G}{\epsilon}\right) + 2G\log n. \tag{A34}$$

Proof. The proof of the upper bounds is similar to the proof of Theorem C.1 in Ref. [154]. We first prove the upper bound for the diamond distance.

Let $\epsilon \in (0, 1]$, and define $\epsilon' = \epsilon/2G$. Then, by Lemma 9, there exists an ϵ'-covering net $\tilde{\mathcal{N}}_{\epsilon'}$ of the set of two-qubit unitaries $U(2^2)$ with respect to the spectral norm $\|\cdot\|$ of size

$$|\tilde{\mathcal{N}}_{\epsilon'}| \leq \left(\frac{6}{\epsilon'}\right)^{32} = \left(\frac{12G}{\epsilon}\right)^{32}. \tag{A35}$$

This bound applies when the two-qubit unitary acts on a fixed set of two qubits. We can consider two-qubit unitaries that act on any of the n qubits. Let $U^{2q} \subset U(2^n)$ denote this set of two-qubit unitaries that can act on any pair of the n qubits of the system. Because there are $\binom{n}{2}$ pairs of qubits that the unitary could act on, the size of the covering net $\tilde{\mathcal{N}}_{\epsilon',n}$ of U^{2q} is bounded by

$$|\tilde{\mathcal{N}}_{\epsilon',n}| \leq \binom{n}{2}\left(\frac{12G}{\epsilon}\right)^{32}. \tag{A36}$$

Recall that we want to find a covering net for the set U^G of n-qubit unitaries consisting of G two-qubit gates. Any unitary $U \in U^G$ can be written as $U_G U_{G-1} \cdots U_1$ for $U_i \in U^{2q}$, where we suppress the tensor product with identity for readability. We consider the set of unitaries obtained by multiplying elements of the covering net $\tilde{\mathcal{N}}_{\epsilon',n}$ of U^{2q}. Namely, we define

$$\mathcal{N}_\epsilon \triangleq \{U_G U_{G-1} \cdots U_1 | U_i \in \tilde{\mathcal{N}}_{\epsilon',n}, 1 \leq i \leq G\}. \tag{A37}$$

Let $U \in U^G$ be any arbitrary unitary that can be implemented by G two-qubit gates, i.e., it can be written as $U = U_G U_{G-1} \cdots U_1$ for $U_i \in U^{2q}$. As $\tilde{\mathcal{N}}_{\epsilon',n}$ is an ϵ'-covering net of the set U^{2q} of two-qubit unitaries, for each U_i comprising the circuit U, we can find a $\tilde{U}_i \in \tilde{\mathcal{N}}_{\epsilon',n}$ such that $\|U_i - \tilde{U}_i\| \leq \epsilon'$ for all $1 \leq i \leq G$, where $\|\cdot\|$ denotes the spectral norm. Then, the unitary $\tilde{U} \triangleq \tilde{U}_G \tilde{U}_{G-1} \cdots \tilde{U}_1 \in \mathcal{N}_{\epsilon_1}$ satisfies

$$\begin{aligned}d_\diamond(U, \tilde{U}) &\leq \sum_{i=1}^{G} d_\diamond(U_i, \tilde{U}_i) \leq 2\sum_{i=1}^{G}\|U_i - \tilde{U}_i\| \\ &\leq 2G\epsilon' = \epsilon,\end{aligned} \tag{A38}$$

where we have employed the subadditivity of the diamond distance (Lemma 2) in the first inequality and then used the relationship between the diamond norm and the spectral norm in the second inequality (Lemma 1). In the last inequality, we have used that $\|U_i - \tilde{U}_i\| \leq \epsilon'$ and $\epsilon' = \epsilon/2G$.

Thus, $\mathcal{N}_\epsilon$ is an ϵ-covering net of the set U^G of n-qubit unitaries that can be implemented by G two-qubit gates with respect to the diamond distance. By the definition of $\mathcal{N}_\epsilon$, we have $|\mathcal{N}_\epsilon| = |\tilde{\mathcal{N}}_{\epsilon',n}|^G$, since each unitary in the length G strings of unitaries comprising elements of $\mathcal{N}_\epsilon$ are chosen from $\tilde{\mathcal{N}}_{\epsilon',n}$. Then,

$$|\mathcal{N}_\epsilon| \leq \binom{n}{2}^G \left(\frac{12G}{\epsilon}\right)^{32G} \leq n^{2G}\left(\frac{12G}{\epsilon}\right)^{32G}. \tag{A39}$$

Taking the logarithm gives the desired result for the diamond distance.

We can argue similarly for the normalized Frobenius distance d_F. Specifically, we make use of the subadditivity

of $\|\cdot\|_F$: $\forall U_1, V_1, U_2, V_2 \in U(2^n)$, we have

$$\begin{aligned}\|U_2U_1 - V_2V_1\|_F &\le \|U_2U_1 - U_2V_1\|_F \\ &\quad + \|U_2V_1 - V_2V_1\|_F \\ &= \|U_1 - V_1\|_F + \|U_2 - V_2\|_F, \end{aligned} \tag{A40}$$

where we have used the triangle inequality and the fact that $\|\cdot\|_F$ is unitary invariant.

Consider any $U \in U^G, U = U_G \cdots U_1$, where $U_i, 1 \le i \le T$ are two-qubit unitaries acting on some pair of qubits. Take $\epsilon' = \epsilon/G$ and let $\mathcal{N}_{\epsilon'}$ be an ϵ'-covering net of $U(2^2)$ with respect to $\|\cdot\|_F$. Then, there exist $V_i \in \mathcal{N}, 1 \le i \le G$ such that $\|U_i - V_i\| \le \epsilon/G$ when the V_i are placed on the corresponding qubits. Let $V = V_G \cdots V_1$. By subadditivity, we have

$$\begin{aligned}\|U - V\|_F &\le \sum_{i=1}^{G} \sqrt{2^{n-2}} \|U_i - V_i\|_F \\ &\le \sqrt{2^{n-2}} G\epsilon' = \sqrt{2^{n-2}}\epsilon, \end{aligned} \tag{A41}$$

where we have used the facts that the Frobenius norm is multiplicative with respect to tensor products and that an $(n-2)$-qubit identity has a Frobenius norm equal to $\sqrt{2^{n-2}}$. Therefore, the set of $V = V_G \cdots V_1$, where $V_i \in U(2^2)$ and acting on all possible pairs of qubits is a $(2^{n-2}\epsilon)$-covering net of U^G. Since the number of choices for qubits to act on is $\binom{n}{2}$ for each V_i, we have

$$\begin{aligned}\mathcal{N}(U^G, \|\cdot\|_F, \sqrt{2^{n-2}}\epsilon) &\le \left[\binom{n}{2}\mathcal{N}(U(2^2), \|\cdot\|_F, \epsilon/G)\right]^G \\ &\le n^{2G}\left(\frac{c_2 G\sqrt{2^2}}{\epsilon}\right)^{16G}, \end{aligned} \tag{A42}$$

where we have used Lemma 9. Redefining ϵ to be $\epsilon/\sqrt{2^2}$ and switching to the normalized d_F, we obtain

$$\log \mathcal{N}(U^G, d_F, \epsilon) \le 16G \log\left(\frac{c_2 G}{\epsilon}\right) + 2G \log n. \tag{A43}$$

Finally, we prove the lower bound. For this, we consider a particular set of circuit structures where all the G gates are placed on the first $k \le n$ qubits. The set of unitaries that can be implemented by such circuits is denoted by $U_G^{\le k} \subseteq U^G$. From the theory of universal quantum gates (see Ref. [146]), we know that to implement an arbitrary k-qubit unitary, we only need $G_k = \mathcal{O}(4^k)$ two-qubit gates that can implement single-qubit gates and a controlled-NOT (CNOT). That is, there exists a universal constant $C > 0$, such that $C4^k \ge G_k$. Therefore, for any integer $k \le n$ satisfying $C4^k \le G$, we have $G \ge G_k$. Then, all possible k-qubit unitaries can be implemented with these G gates: $U^n(2^k) = \{U \otimes I_{2^{n-k}} : U \in U(2^k)\} \subseteq U_G^{\le k} \subseteq U^G$, where $U^n(2^k)$ denotes the set obtained by embedding the k-qubit unitaries into the n-qubit unitaries via tensor multiplication with the identity. Thus $\mathcal{N}(U^G, \|\cdot\|_F, \epsilon) \ge \mathcal{N}(U^n(2^k), \|\cdot\|_F, 2\epsilon)$ by monotonicity.

Next, we prove that $\mathcal{N}(U^n(2^k), \|\cdot\|_F, 2\epsilon) \ge \mathcal{N}(U(2^k), \|\cdot\|_F, 2\epsilon/\sqrt{2^{n-k}})$. To do this, we take a minimal 2ϵ-covering net $\mathcal{N}$ of $U^n(2^k)$ with $|\mathcal{N}| = \mathcal{N}(U^n(2^k), \|\cdot\|_F, 2\epsilon)$. Hence, $\forall U \in U(2^k)$, $U \otimes I_{2^{n-k}} \in U^n(2^k), \exists V \otimes I_{2^{n-k}} \in \mathcal{N}$, such that $\|U - V\|_F = \|U \otimes I_{2^{n-k}} - V \otimes I_{2^{n-k}}\|_F/\sqrt{2^{n-k}} \le 2\epsilon/\sqrt{2^{n-k}}$. Therefore, $\{V : V \otimes I_{2^{n-k}} \in \mathcal{N}\}$ forms a $2\epsilon/\sqrt{2^{n-k}}$-covering net of $U(2^k)$ and we have $\mathcal{N}(U^n(2^k), \|\cdot\|_F, 2\epsilon) \ge \mathcal{N}(U(2^k), \|\cdot\|_F, 2\epsilon/\sqrt{2^{n-k}})$.

Combining the above inequalities, we have

$$\begin{aligned}\log \mathcal{N}(U^G, \|\cdot\|_F, \epsilon) &\ge \log \mathcal{N}(U^n(2^k), \|\cdot\|_F, 2\epsilon) \\ &\ge \log \mathcal{N}(U(2^k), \|\cdot\|_F, 2\epsilon/\sqrt{2^{n-k}}) \\ &\ge 2^{2k} \log \frac{c_1\sqrt{2^n}}{2\epsilon}, \end{aligned} \tag{A44}$$

where the last inequalities follow from Lemma 9. The largest possible k is given by $k = \lfloor \log_4(G/C) \rfloor \ge \log_4 G/(4C)$. Thus, by redefining ϵ to be $\epsilon/\sqrt{2^n}$ and switching to d_F, we arrive at

$$\log \mathcal{N}(U^G, d_F, \epsilon) \ge \frac{G}{4C} \log \frac{c_1}{2\epsilon}. \tag{A45}$$

This completes the proof of Theorem 8. ■

The d_F covering-number bounds in Theorem 8 do not yet properly take into account the global $U(1)$ phase. To obtain the covering number for the average-case distance d_{avg}, which is equivalent to the quotient-normalized Frobenius distance d_F' [Lemma 4, item (1)], we need to quotient out the global phase. This is formalized in the following lemma.

Lemma 10 (Packing number of quotient distance metric; variant of Ref. [158, Lemma 4]). For any d-dimensional unitaries U and V, let $d_F(U, V) = \|U - V\|_F/\sqrt{d}$ be the normalized Frobenius distance and let $d_F'(U, V) = \min_{W \in U(1)} d_F(U, VW)$ be the corresponding quotient distance. Then, there exists a universal constant $c_2 > 0$ such that the packing number of any set $\mathcal{U} \subseteq U(d)$ with respect to d_F and d_F' satisfies

$$\begin{aligned}&\log \mathcal{M}(\mathcal{U}, d_F, 4\epsilon) - \log(c_2/\epsilon) \\ &\quad \le \log \mathcal{M}(\mathcal{U}, d_F', \epsilon) \le \log \mathcal{M}(\mathcal{U}, d_F, \epsilon). \end{aligned} \tag{A46}$$

Proof. We focus on the lower bound first. Take a minimal ϵ-covering $\mathcal{N}_1$ of $\mathcal{U}$ with respect to d_F' and a minimal ϵ-covering $\mathcal{N}_2$ of $U(1)$ with respect to the absolute-value

distance $d_A(e^{i\phi}, e^{-i\phi'}) = |e^{i\phi} - e^{-i\phi'}|$. Then, for any $U \in \mathcal{U}$, there exists $V \in \mathcal{N}_1$ such that $d'_F(U, V) \leq \epsilon$. Let $e^{i\phi^\star} = arg\,min_{e^{i\phi}} d_F(U, Ve^{i\phi})$. Then, $d_F(U, Ve^{i\phi^\star}) \leq \epsilon$ and there exists $e^{i\phi'} \in \mathcal{N}_2$ such that $d_A(e^{i\phi^\star}, e^{i\phi'}) \leq \epsilon$. Therefore,

$$\begin{aligned} d_F(U, Ve^{i\phi'}) &\leq d_F(U, Ve^{i\phi^\star}) + d_F(Ve^{i\phi^\star}, Ve^{i\phi'}) \\ &= d_F(U, Ve^{i\phi^\star}) + d_F(Ie^{i\phi^\star}, Ie^{i\phi'}) \leq 2\epsilon, \end{aligned} \tag{A47}$$

where we have used the triangle inequality, d_F being unitary invariant, and $d_F(Ie^{i\phi^\star}, Ie^{i\phi'}) = (1/\sqrt{d})\|Ie^{i\phi^\star} - Ie^{i\phi'}\|_F = (\|I\|_F/\sqrt{d})|e^{i\phi^\star} - e^{i\phi'}| = d_A(e^{i\phi^\star}, e^{i\phi'}) \leq \epsilon$. Hence, the set $\{Ve^{i\phi'} : V \in \mathcal{N}_1, e^{i\phi'} \in \mathcal{N}_2\}$ is a (2ϵ)-covering of $\mathcal{U}$ with respect to d_F. Then,

$$\mathcal{N}(\mathcal{U}, d_F, 2\epsilon) \leq \mathcal{N}(\mathcal{U}, d'_F, \epsilon)\mathcal{N}(U(1), d_A, \epsilon). \tag{A48}$$

Therefore, using the equivalence of covering and packing (Lemma 7) and the covering-number bound for $U(1)$ (Lemma 9), we arrive at

$$\log \mathcal{M}(\mathcal{U}, d'_F, \epsilon) \geq \log \mathcal{M}(\mathcal{U}, d'_F, 4\epsilon) - \log(c_2/\epsilon). \tag{A49}$$

For the upper bound, note that $\forall U, V \in \mathcal{U}$, we have

$$d'_F(U, V) = \min_{e^{i\phi} \in U(1)} d_F(U, Ve^{i\phi}) \leq d_F(U, V). \tag{A50}$$

Therefore, a maximal ϵ-packing net with respect to d'_F is an ϵ-packing net with respect to d_F. Therefore,

$$\mathcal{M}(\mathcal{U}, d'_F, \epsilon) \leq \mathcal{M}(\mathcal{U}, d_F, \epsilon). \tag{A51}$$

This concludes the proof of Lemma 10. ■

With Lemma 10, we can obtain the covering number of G-gate unitaries with respect to the average-case distance.

Corollary 1 (Covering number with average-case distance). Let $U^G \subseteq U(2^n)$ be the set of n-qubit unitaries that can be implemented by G two-qubit gates. Then, for any $\epsilon \in (0, 1]$, there exist universal and positive constants c_1, c_2, and C such that for $1 \leq G/C \leq 4^{n+1}$, the metric entropy of U^G with respect to the average-case distance $d_{\text{avg}}(U, V) = \sqrt{\mathbb{E}_{|\psi\rangle}[d_{\text{tr}}(U|\psi\rangle, V|\psi\rangle)^2]}$, where the expectation value is over the Haar measure, can be bounded as

$$\begin{aligned} \frac{G}{4C}\log\left(\frac{c_1}{8\epsilon}\right) - \log\left(\frac{c_2}{2\epsilon}\right) &\leq \log \mathcal{N}(U^G, d_{\text{avg}}, \epsilon) \\ &\leq 16G\log\left(\frac{c_2 G}{\epsilon}\right) + 2G\log n. \end{aligned} \tag{A52}$$

Proof. The corollary follows directly from Theorem 8, Lemma 10, and the equivalence of d'_F and d_{avg} [Lemma 4, item (1)]. ■

3. Classical shadows and hypothesis selection

Our proofs of the sample-complexity upper bounds crucially rely on a known algorithm for quantum hypothesis selection [160]. The high-level idea is to find a covering net over all unitaries consisting of only G two-qubit gates and to then use quantum hypothesis selection to identify a candidate in the covering net close to the unknown target state or unitary. A similar idea has previously appeared in Ref. [113]. In this section, we discuss the quantum hypothesis-selection algorithm from Ref. [160] and prove a performance guarantee when basing it on classical shadow tomography [110].

The quantum hypothesis-selection algorithm takes as input (classical descriptions of) a set of hypothesis states $\sigma_1, \ldots, \sigma_m$ and quantum copies of an unknown state ρ. Using these copies, the algorithm identifies a hypothesis state σ_k that is close to the unknown state ρ in trace distance. Importantly, the quantum hypothesis-selection black box reduces to shadow tomography [42], i.e., one can use the shadow-tomography protocol as a black box to solve quantum hypothesis selection. To obtain a better sample-complexity scaling, we instead utilize classical shadow tomography [110].

Recall that a classical shadow is a succinct classical description of a quantum state that allows us to predict many expectation values accurately. One can construct this classical shadow description by applying a random unitary to the quantum state and measuring in the computational basis. The most prevalent examples are random Clifford measurements, where the random unitary is chosen to be a random Clifford circuit, or random Pauli measurements, where the random unitary is chosen to be a tensor product of random Pauli gates. Moreover, we have the following rigorous guarantee for using classical shadows to predict expectation values.

Theorem 9 (Theorem 1 in Ref. [110]). Let $O_1, \ldots, O_M$ be Hermitian $2^n \times 2^n$ matrices and let $\epsilon, \delta \in [0, 1]$. Then,

$$N = \mathcal{O}\left(\frac{\log(M/\delta)}{\epsilon^2} \max_{1 \leq i \leq M} \left\| O_i - \frac{\text{tr}(O_i)}{2^n}\mathbb{I} \right\|^2_{\text{shadow}}\right) \tag{A53}$$

copies of an unknown quantum state ρ suffice to predict $\hat{o}_i$ such that

$$|\hat{o}_i - \text{tr}(O_i\rho)| \leq \epsilon \tag{A54}$$

for all $1 \leq i \leq M$, with probability at least $1 - \delta$.

Here, $\|\cdot\|_{\text{shadow}}$ denotes the shadow norm, which depends on the ensemble of unitary transformations used to create the classical shadow. For instance, in the case of random Cliffords, the shadow norm can be controlled via the (unnormalized) Frobenius norm (cf. [110, Proposition S1]).

Now, we can prove a new guarantee for the quantum hypothesis selection by replacing the shadow tomography with the classical shadow in the proof in Ref. [160].

Proposition 1 (Proposition 5.3 in Ref. [160]; classical-shadow version). Let $0 < \epsilon, \delta < 1/2$. Given access to unentangled copies of a pure quantum state ρ and classical descriptions of m fixed pure hypothesis states $\sigma_1, \ldots, \sigma_m$, there exists a quantum algorithm that selects σ_k such that $d_{\mathrm{tr}}(\rho, \sigma_k) \leq 3\eta + \epsilon$ with probability at least $1 - \delta$, where $\eta = \min_i d_{\mathrm{tr}}(\rho, \sigma_i)$. Moreover, this algorithm uses

$$N = \mathcal{O}\left(\frac{\log(m/\delta)}{\epsilon^2}\right) \tag{A55}$$

copies of the quantum state ρ.

In Ref. [160], the authors prove the guarantee on the quantum hypothesis-selection algorithm using Helstrom's theorem. We follow a similar proof. Thus, we first state Helstrom's theorem and recall a corollary of it, which will be useful in the proof of Proposition 1.

Theorem 10 (Helstrom's theorem [142]). Consider two d-dimensional quantum states ρ and σ. Then, the trace distance between ρ and σ can be written as

$$\frac{1}{2}\|\rho - \sigma\|_1 = \max_{\|O\|_\infty \leq 1} |\mathrm{tr}(O\rho) - \mathrm{tr}(O\sigma)|, \tag{A56}$$

where the maximum is taken over all observables $O \in \mathbb{C}^{d\times d}$.

Corollary 2. Consider two d-dimensional quantum states, ρ and σ. Then, there exists an observable A achieving the maximum such that

$$\mathrm{tr}(A\rho) - \mathrm{tr}(A\sigma) = \frac{1}{2}\|\rho - \sigma\|_1. \tag{A57}$$

Proof of Corollary 2. We will construct an observable A that maximizes $\mathrm{tr}(O(\rho - \sigma))$ over all observables O with $\|O\|_\infty \leq 1$. Choose a representation of $\rho - \sigma$ in terms of eigenstates $|v\rangle$. Suppose that the eigenvalues are discrete:

$$(\rho - \sigma)\,|v\rangle = \lambda_v\,|v\rangle\,. \tag{A58}$$

Then, we can write the quantity that we wish to maximize as

$$\mathrm{tr}(O(\rho - \sigma)) = \sum_v \lambda_v \langle v|O|v\rangle. \tag{A59}$$

We can maximize this by choosing A such that

$$\langle v|A|v\rangle = \begin{cases} 1, & \text{if } \lambda_v > 0, \\ 0, & \text{if } \lambda_v \leq 0. \end{cases} \tag{A60}$$

In this way, we can write A as a sum of projectors

$$A = \sum_{v:\lambda_v > 0} |v\rangle\langle v|\,. \tag{A61}$$

This maximizes $\mathrm{tr}(O(\rho - \sigma))$, so the corollary has been proven. ■

With this, we can now prove Proposition 1.

Proof of Proposition 1. The proof of Proposition 5.3 in Ref. [160] uses shadow tomography as a black box. We follow the same strategy but use classical shadow tomography [110] instead of shadow tomography. Recall that in Ref. [160], the authors run the shadow-tomography algorithm from Ref. [42], with observables given by Helstrom's theorem [142]. This is the key step that uses samples of the unknown quantum state ρ, so we need to analyze it when using classical shadow instead of shadow tomography. In our setting, Corollary 2 states that for any $i \neq j$, there exists an observable A_{ij} such that

$$\mathrm{tr}(A_{ij}\sigma_i) - \mathrm{tr}(A_{ij}\sigma_j) = \frac{1}{2}\|\sigma_i - \sigma_j\|_1. \tag{A62}$$

Thus, the algorithm in Ref. [160] uses $M = \binom{m}{2} = \mathcal{O}(m^2)$ observables $\{A_{ij}\}$ to select the hypothesis state, where m is the size of the hypothesis set. Using classical shadow instead of shadow tomography requires

$$N = \tilde{\mathcal{O}}\left(\frac{\log(M/\delta)}{\epsilon^2} \max_{i,j} \left\|A_{ij} - \frac{\mathrm{tr}(A_{ij})}{2^n}\mathbb{I}\right\|^2_{\mathrm{shadow}}\right) \tag{A63}$$

copies of ρ by Theorem 9, where M is the number of observables A_{ij} that we want to predict. Here, $M = \mathcal{O}(m^2)$, so that we require

$$N = \tilde{\mathcal{O}}\left(\frac{\log(m/\delta)}{\epsilon^2} \max_{i,j} \left\|A_{ij} - \frac{\mathrm{tr}(A_{ij})}{2^n}\mathbb{I}\right\|^2_{\mathrm{shadow}}\right) \tag{A64}$$

copies of ρ. We claim that

$$\max_{i,j} \left\|A_{ij} - \frac{\mathrm{tr}(A_{ij})}{2^n}\mathbb{I}\right\|^2_{\mathrm{shadow}} = \mathcal{O}(1). \tag{A65}$$

The lemma then follows from this claim. We can prove this bound on the shadow norm using the construction of the observables A_{ij} from Helstrom's theorem, as seen in Corollary 2. In our case, the states σ_i are pure and hence of rank 1. Thus, the rank of $\sigma_i - \sigma_j$ is at most 2, so that A_{ij} is a projector of rank at most 2. Thus, the Frobenius norm of every A_{ij} is $\mathcal{O}(1)$ and, by Ref. [110, Proposition S1], the same holds for the shadow norm of the centered version of A_{ij}. ■

4. Characterizing the complexity of function classes

In the proof of Theorem 7 (in Appendix D), we will need to characterize the complexity of certain function classes. The following definitions will be useful. Throughout the work, we use $\mathcal{Y}^{\mathcal{X}}$ to denote the set of functions from $\mathcal{X}$ to $\mathcal{Y}$.

Definition 4 (Growth function [161]). Let $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a class of functions with finite target space $\mathcal{Y}$. For every subset $\Xi \subseteq X$, define the restriction of $\mathcal{F}$ to Ξ as $\mathcal{F}|_{\Xi} = \{f \in \mathcal{Y}^{\Xi} : \exists F \in \mathcal{F}, \forall x \in \Xi, f(x) = F(x)\}$. We define the growth function Γ of $\mathcal{F}$ as

$$\Gamma(\mu) = \max_{\Xi \subseteq \mathcal{X}: |\Xi| \leq \mu} |\mathcal{F}|_{\Xi}| \tag{A66}$$

for any $\mu \in \mathbb{N}$.

The growth function characterizes the size of $\mathcal{F}$ when restricted to a domain of μ points. With the growth function, we can define the VC dimension that characterizes the complexity of binary functions.

Definition 5 (VC dimension [161]). The Vapnik-Chervonenkis (VC) dimension of a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ is defined as

$$\mathrm{VCdim}(\mathcal{F}) = \max\{\mu \in \mathbb{N} : \Gamma(\mu) = 2^{\mu}\}, \tag{A67}$$

or ∞ if the maximum does not exist. Here, $\Gamma(\mu)$ is the growth function of $\mathcal{F}$. Or, equivalently, $\mathrm{VCdim}(\mathcal{F})$ is the largest $D \in \mathbb{N} \cup \{\infty\}$ such that there exists a set of points $\{x_i\}_{i=1}^{D} \subseteq \mathcal{X}$ such that for all $C \subseteq [D]$, there is a function $f \in \mathcal{F}$ satisfying

$$f(x_i) = 1 \iff i \in C. \tag{A68}$$

These points are said to be shattered by $\mathcal{F}$.

To go beyond binary functions, we can use the pseudodimension defined below.

Definition 6 (Pseudodimension [75]). The pseudodimension of a real-valued function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as

$$\begin{aligned}\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\{&\mathcal{X} \times \mathbb{R} \ni (x,y) \\ &\to \mathrm{sgn}[f(x) - y] : f \in \mathcal{F}\}).\end{aligned} \tag{A69}$$

Or, equivalently, $\mathrm{Pdim}(\mathcal{F})$ is the largest $D \in \mathbb{N} \cup \{\infty\}$ such that there exists a set of points $\{(x_i, y_i)\}_{i=1}^{D} \subseteq \mathcal{X} \times \mathbb{R}$ such that for all $C \subseteq [D]$, there is a function $f \in \mathcal{F}$ satisfying

$$f(x_i) \geq y_i \iff i \in C. \tag{A70}$$

These points are said to be pseudoshattered by $\mathcal{F}$.

We will also use the fat-shattering dimension, a scale-sensitive variant of the pseudodimension.

Definition 7 (Fat-shattering dimension [76]). Let $\alpha > 0$. The α-fat-shattering dimension $\mathrm{fat}(\mathcal{F}, \alpha)$ of a real-valued function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as the largest $D \in \mathbb{N} \cup \{\infty\}$ such that there exists a set of points $\{(x_i, y_i)\}_{i=1}^{D} \subseteq \mathcal{X} \times \mathbb{R}$ such that for all $C \subseteq [D]$, there is a function $f \in \mathcal{F}$ satisfying

$$\begin{aligned} f(x_i) &\geq y_i + \alpha \quad \text{if } i \in C, \\ f(x_i) &\leq y_i - \alpha \quad \text{if } i \notin C. \end{aligned} \tag{A71}$$

Such a set of points is said to be α-fat-shattered by $\mathcal{F}$.

5. Cryptography

Our computational-complexity lower bounds rely on cryptographic primitives such as pseudorandom functions [148] and pseudorandom quantum states [147,149]. A family of pseudorandom functions is a set of functions such that sampling from this family is indistinguishable from a uniformly random function. We present the formal definition below, following the presentation in Ref. [45].

Definition 8 (Pseudorandom functions (PRFs) [148]). Let λ denote the security parameter. Let $\mathcal{K} = \{\mathcal{K}_\lambda\}_{\lambda \in \mathbb{N}}$ be an efficiently sampleable key space. Let $\mathcal{X} = \{\mathcal{X}_\lambda\}_{\lambda \in \mathbb{N}}, \{\mathcal{Y}_\lambda\}_{\lambda \in \mathbb{N}}$ be collections of finite sets. Let $\mathcal{F} = \{f_\lambda\}_{\lambda \in \mathbb{N}}$ be a family of efficiently computable keyed functions $f_\lambda : \mathcal{K}_\lambda \times \mathcal{X}_\lambda \to \mathcal{Y}_\lambda$. $\mathcal{F}$ is a *pseudorandom function* if for every polynomial-time probabilistic algorithm Adv, there exists a negligible function $\mathrm{negl}(\cdot)$ such that for every security parameter $\lambda \in \mathbb{N}$,

$$\left| \Pr_{\mathbf{k} \leftarrow \mathcal{K}_\lambda} [\mathrm{Adv}^{f_\lambda(\mathbf{k}, \cdot)}(\cdot) = 1] - \Pr_{g \in \mathcal{U}_\lambda} [\mathrm{Adv}^{g}(\cdot) = 1] \right| \leq \mathrm{negl}(\lambda), \tag{A72}$$

where the key $\mathbf{k}$ is picked uniformly at random from the key space $\mathcal{K}_\lambda$ and g is picked uniformly at random from $\mathcal{U}_\lambda$, the set of all functions from $\mathcal{X}_\lambda$ to $\mathcal{Y}_\lambda$. Here, $\mathrm{negl}(\lambda)$ denotes a negligible function, i.e., a function that grows more slowly than any inverse polynomial in λ.

Concretely, it is common to take the input and output spaces to be $\mathcal{X}_\lambda = \{0,1\}^m$ and $\mathcal{Y}_\lambda = \{0,1\}$ for some input length $m = m(\lambda)$ that depends on the security parameter λ. We consider this setting throughout the work.

Definition 9 (Quantum secure PRFs [45]). Let λ denote the security parameter. A pseudorandom function is *quantum secure against* $t(\lambda)$ *adversaries* if it satisfies Definition 8, where Adv is a $t(\lambda)$-time quantum algorithm with quantum query access to $f_{\mathbf{k}}$ and g. When $t(\lambda) = \mathrm{poly}(\lambda)$, we say that the PRF is *quantum secure*.

There are several constructions for implementing PRFs with low-depth circuits [45,162,163]. We will focus on the construction of Ref. [45], which relies on the assumption that the RingLWE problem [43] is hard even for quantum computers. Specifically, we assume that RingLWE cannot be solved by a quantum computer in subexponential time, which is a commonly believed cryptographic assumption [44–48]. Here, RingLWE is a variant of the more well-known "learning with errors" problem [44] over polynomial rings. The RingLWE problem is to find a secret ring element $s \in R_q \triangleq \mathbb{Z}_q[x]/\langle x^\lambda - 1\rangle$ given pairs $(a, a \cdot s + e \bmod R_q)$, where λ denotes the security parameter, e is some error, and q is a parameter of the problem. We only state this informally here and refer the reader to Ref. [43] for a formal definition and discussion. In Ref. [45], assuming that RingLWE cannot be solved by quantum computers in $t(\lambda)$ time, the construction in Ref. [45] produces a PRF secure against $\mathcal{O}(t(\lambda))$ quantum adversaries that is implementable by constant-depth polynomial-size circuits. We state the precise result below.

Theorem 11 (Lemmas 3.15 and 3.16 in Ref. [45]). Let λ denote the security parameter. Let the input size be $m = m(\lambda) = \omega(\log\lambda)$ and set the parameter $q = \lambda^{\omega(1)}$ to be a power of 2 such that $\log(q) \le \mathcal{O}(\mathrm{poly}(\lambda))$. Let $\mathcal{K} = \{\mathcal{K}_\lambda\}_{\lambda\in\mathbb{N}}$, where $\mathcal{K}_\lambda = R_q^{m+1}$. There exists a PRF $\mathcal{RF} = \{f_\lambda\}_{\lambda\in\mathbb{N}}$, where $f_\lambda : R_q^{m+1} \times \{0,1\}^m \to \{0,1\}$, satisfying the following two properties:

(1) Every $f_\lambda(\mathbf{k},\cdot) \in \mathcal{RF}$ with $\mathbf{k} \in \mathcal{K}_\lambda$ can be computed by a TC^0 circuit.
(2) Suppose there exists a distinguisher $\mathcal{D}$ for $\mathcal{RF}$, i.e., there exists an $\mathcal{O}(t(\lambda))$-time quantum algorithm $\mathcal{D}$ that satisfies

$$\left|\Pr_{\mathbf{k}\leftarrow\mathcal{K}_\lambda}[\mathcal{D}^{|f_\lambda(\mathbf{k},\cdot)\rangle}(\cdot)=1] - \Pr_{g\in\mathcal{U}}[\mathcal{D}^{|g\rangle}(\cdot)=1]\right| > \mathrm{negl}(\lambda), \tag{A73}$$

where the key $\mathbf{k}$ is picked uniformly at random from the key space $\mathcal{K}_\lambda$, g is picked uniformly at random from $\mathcal{U}$, the set of all functions from $\mathcal{X}_\lambda$ to $\mathcal{Y}_\lambda$, and $\mathcal{D}^{|f_\lambda(\mathbf{k},\cdot)\rangle}$ indicates that $\mathcal{D}$ has quantum oracle access to the function $f_\lambda(\mathbf{k},\cdot)$. Then, there exists a $t(\lambda)$-time quantum algorithm that solves RingLWE.

In property (2), this is equivalent to saying that the PRF is quantum secure against $\mathcal{O}(t(\lambda))$ adversaries, assuming that RingLWE cannot be solved by a $t(\lambda)$-time quantum algorithm. Also, note that in property (1), TC^0 circuits refer to constant-depth polynomial-size circuits with unbounded fan-in AND, OR, NOT, and MAJORITY gates. We claim that every TC^0 circuit has a quantum circuit computing the same function with polylogarithmic overhead in depth.

Proposition 2 (Quantum circuits for TC^0). Let C be a TC^0 circuit on m inputs computing some Boolean function $f : \{0,1\}^m \to \{0,1\}$. Then, there exists a quantum circuit C' on $n = \mathcal{O}(\mathrm{poly}(m))$ qubits of size $\mathcal{O}(n\mathrm{polylog}(n))$ and depth $\mathcal{O}(\mathrm{polylog}(n))$ that implements f.

Here, when we say that C' implements the function f, we mean that $C'|x\rangle|z\rangle = C'|x\rangle|z \oplus f(x)\rangle$.

Proof. Note that the number of qubits is $n = \mathcal{O}(\mathrm{poly}(m))$, because after each gate in the classical circuit C, we must store the result in an ancilla qubit to maintain unitarity. Recall that TC^0 circuits are constant-depth polynomial-size circuits with unbounded fan-in AND, OR, NOT, and MAJORITY gates. Thus, it suffices to find the depth of implementing each of these gates quantumly. The size then follows because a circuit of depth d on n qubits can have at most nd gates. NOT gates can clearly be implemented in constant depth, since this is just an X gate. An AND gate with m inputs can be completed in logarithmic depth by computing AND pairwise with the controlled-NOT (CNOT). Similarly, we can compute an OR gate with the same logarithmic depth. It remains to analyze the depth needed for computing a MAJORITY gate. Recall that the MAJORITY gate is defined as

$$\mathrm{MAJORITY}(x_1,\ldots,x_m) = \left\lfloor \frac{1}{2} + \frac{\left(\sum_{i=1}^m x_i\right) - 1/2}{m} \right\rfloor = \left\lfloor \frac{1}{2} + \frac{\sum_{i=1}^m x_i}{m} - \frac{1}{2m} \right\rfloor. \tag{A74}$$

Here, addition is done over the integers and $x_i \in \{0,1\}$. We first analyze the depth or size required for the addition $\sum_{i=1}^m x_i$. Note that the maximum value of this sum is m, which can be stored in $\mathcal{O}(\log m)$ bits. Thus, we can write each of the x_i in binary using $\log m$ bits by padding with zeros and perform addition in this way. We can perform the addition of the m inputs pairwise, parallelized to $\mathcal{O}(\log m)$ depth and requiring $\mathcal{O}(m)$ addition operations. Moreover, one can perform these addition operations using quantum circuits of size and depth $\mathcal{O}(\log m)$ [164]. The construction in Ref. [164] uses Toffoli gates but these can be decomposed into two-qubit gates with constant overhead [165]. In total, we have that $\sum_{i=1}^m x_i$ can be implemented by a quantum circuit of depth $\mathcal{O}(\log^2 m)$.

To divide this sum by m, note that there exist classical Boolean circuits for integer division of depth $\mathcal{O}(\log\log m)$, since our inputs can be represented in binary using $\log m$ bits [166]. These Boolean circuits use only standard AND, OR, and NOT gates. As explained previously, these can be implemented quantumly and for fan-in-2 AND and OR gates, this can be done with constant overhead. Thus, this division step requires depth $\mathcal{O}(\log\log m)$ in total.

Finally, we need to compute the remaining addition or subtraction and floor operations. The addition or subtraction can be ignored since they only occur once, so that the depth is dominated by the other additions. For the floor, because the quantity inside can only be less than or equal to 1, then this is the same as deciding whether or not the quantity inside is less than 1. This can be done in a constant number of operations.

Putting everything together, we see that the circuit depth for implementing a MAJORITY gate is dominated by $\mathcal{O}(\log^2 m)$.

Recall again that TC^0 describes constant-depth polynomial-size circuits with unbounded fan-in AND, OR, NOT, and MAJORITY gates. We have just analyzed the depth for each of these gates individually. In summary, we have computed that $\mathcal{O}(\log m)$ quantum depth is sufficient for AND and OR. Constant $\mathcal{O}(1)$ depth is sufficient for NOT. Finally, $\mathcal{O}(\log^2 m)$ depth is sufficient for MAJORITY. In the overall circuit, this totals to $\mathcal{O}(\mathrm{polylog}(m))$ depth. Because a circuit of depth d on n qubits can have at most nd gates, then the size of this circuit is $\mathcal{O}(n\mathrm{polylog}(m))$ gates. Then, because $n = \mathcal{O}(\mathrm{poly}(m))$, we obtain the claim. ■

Alternatively, we note that one can obtain a similar result using Ref. [167]. As a simple corollary of this along with Theorem 11, we can bound the depth or size of a quantum circuit for computing a PRF.

Corollary 3. Let $\lambda = n$ denote the security parameter. Assuming that RingLWE cannot be solved in $t(n)$ time by a quantum computer, there exists a PRF $\mathcal{F} = \{f_\lambda\}_{\lambda\in\mathbb{N}}$ that is secure against $\mathcal{O}(t(n))$ quantum adversaries such that for keys $\mathbf{k} \in \mathcal{K}_\lambda$ (for the same key space as in Theorem 11), $f_\lambda(\mathbf{k}, \cdot) : \{0,1\}^m \to \{0,1\}$ is computable by an n-qubit quantum circuit of size $\mathcal{O}(n\mathrm{polylog}(n))$ and depth $\mathcal{O}(\mathrm{polylog}(n))$.

Note here that by the above analysis, we have $m = \omega(\log(\lambda))$. Since $\mathcal{O}(\mathrm{poly}(m))$ qubits suffice to implement these PRFs, we can take $n = \lambda$, similar to Ref. [149].

Our proofs also require the notion of pseudorandom quantum states. Informally, pseudorandom quantum states are ensembles of quantum states that are indistinguishable from Haar-random states to any efficient (quantum) algorithm. Moreover, it is known how to construct these states using efficient quantum circuits. Recently, pseudorandom quantum states have been of great interest in quantum cryptography [168–170] and complexity theory [171]. We define them formally below, following the presentation in Refs. [147,149].

Definition 10 (Pseudorandom quantum states (PRSs) [147]). Let $\lambda = n$ denote the security parameter. Let $\mathcal{K} = \{\mathcal{K}_\lambda\}_{\lambda\in\mathbb{N}}$ be the key space. A keyed family of pure quantum states $\{|\phi_k\rangle\}_{k\in\mathcal{K}_\lambda}$ is *pseudorandom against $t(n)$ adversaries* if the following two conditions hold:

(1) (Efficient generation.) There is a polynomial-time quantum algorithm Gen that generates state $|\phi_k\rangle$ on input k. That is, for all $\lambda \in \mathbb{N}$ and for all $k \in \mathcal{K}_\lambda$, $\mathrm{Gen}(1^\lambda, k) = |\phi_k\rangle$.

(2) (Pseudorandomness.) Any polynomially many copies of $|\phi_k\rangle$ with the same random $k \in \mathcal{K}_\lambda$ are computationally indistinguishable from the same number of copies of a Haar-random state. More precisely, for any $t(n)$-time quantum algorithm $\mathcal{D}$ and any $N = \mathrm{poly}(\lambda)$, there exists a negligible function $\mathrm{negl}(\cdot)$ such that for all $\lambda \in \mathbb{N}$,

$$\left|\Pr_{k\leftarrow\mathcal{K}_\lambda}\left[\mathcal{D}\left(|\phi_k\rangle^{\otimes N}\right)=1\right]-\Pr_{|\psi\rangle\leftarrow\mu}\left[\mathcal{D}\left(|\psi\rangle^{\otimes N}\right)=1\right]\right| \leq \mathrm{negl}(\lambda), \tag{A75}$$

where μ is the Haar measure over pure states on n qubits.

When $t(n) = \mathrm{poly}(n)$, we simply say that the states are *pseudorandom*.

There exist efficient procedures to generate pseudorandom quantum states under standard cryptographic assumptions. In particular, we consider the construction in Ref. [149], which assumes the existence of quantum secure pseudorandom functions.

Proposition 3 (Corollary of Claims 3 and 4 in Ref. [149]). Let $\lambda = n$ denote the security parameter and let $t(n) \geq \mathrm{poly}(n)$. Assuming that RingLWE cannot be solved by a quantum computer in $t(n)$ time, pseudorandom quantum states secure against $\mathcal{O}(t(n))$ adversaries with key space $\mathcal{K}$ (for the same key space as in Theorem 11) can be prepared using n-qubit quantum circuits of depth $\mathcal{O}(\mathrm{polylog}(n))$ and size $\mathcal{O}(n\mathrm{polylog}(n))$.

Proof. Note that using the PRF from Ref. [45] and tracing through the proof of Claim 3 in Ref. [149], one can clearly see that security holds for $\mathcal{O}(t(n))$ adversaries rather than only efficient adversaries. We need to prove that the size and depth are as stated for the construction of pseudorandom quantum states in Ref. [149] using the PRF from Ref. [45]. To obtain the depth and size bounds, we analyze the construction in Ref. [149]. In Claim 3 of Ref. [149], the authors show that their constructed states can be prepared by applying a single layer of Hadamard gates followed by applying a quantum secure PRF. First, the layer of Hadamards has depth 1 and size n. Using the construction from Corollary 3, applying the PRF can then be implemented in $\mathcal{O}(\mathrm{polylog}(n))$ depth and $\mathcal{O}(n\mathrm{polylog}(n))$ size. Thus, overall, the depth and size are dominated by the cost of evaluating the PRF. Moreover, in Claim 4 of Ref. [149], the authors prove that this indeed constructs a pseudorandom quantum state. ■

Note again that the number of qubits n in the quantum circuit depends on the security parameter λ. In fact, due to the construction used, the n depends on λ in the same way as for the PRF construction. Also note that the above PRS-PRF constructions can be implemented using a number of Clifford and T gates of the same order. This is because the TC^0 circuits in the PRF constructions are classical circuits that can be implemented exactly by Toffoli gates, and Toffoli gates can be constructed using a constant number of Clifford and T gates. Also in the PRS construction, the remaining gates are Hadamard gates, which are Clifford gates. Therefore, the computational hardness results in Appendices B 3 and C 5 also apply to Clifford+T circuits of the same gate complexity.

APPENDIX B: LEARNING QUANTUM STATES

Recall that, given copies of a pure state of bounded circuit complexity, we wish to find a classical description for a quantum circuit that approximately implements this state. It is natural to require the learner to output a circuit description, since this ensures that the output of the learner can indeed be used to prepare (approximate) copies of the unknown state. This model is similar in spirit to learning an (approximate) generator for an unknown classical probability distribution [172]. Nevertheless, our sample-complexity results hold for learning classical descriptions beyond circuit descriptions and our computational-complexity results immediately extend to learners that output classical descriptions from which a circuit description can be derived efficiently (e.g., matrix product states and/or operators with constant bond dimension [39,40], stabilizer descriptions, etc.).

Specifically, let $|\psi\rangle = U|0\rangle^{\otimes n}$, where U is a unitary consisting of G two-qubit gates. Throughout this appendix, we denote $\rho \triangleq |\psi\rangle\langle\psi|$. Suppose that we are given N identically prepared copies of ρ. The goal is to learn a classical circuit description of a quantum state $\hat{\rho}$ that is ϵ-close to ρ in trace distance, i.e., $d_{\mathrm{tr}}(\hat{\rho}, \rho) = \|\hat{\rho} - \rho\|_1/2 \leq \epsilon$.

In this appendix, we provide a proof of Theorem 1, which characterizes the sample complexity for this task. We restate the theorem below.

Theorem 12 (State learning; detailed restatement of Theorem 1). Let $\epsilon, \delta > 0$. Suppose that we are given N copies of a pure n-qubit state density matrix $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = U|0\rangle^{\otimes n}$ is generated by a unitary U consisting of G two-qubit gates. Then, any algorithm that can output $\hat{\rho}$ such that $d_{\mathrm{tr}}(\hat{\rho}, \rho) \leq \epsilon$ with probability at least $1 - \delta$ requires at least

$$N = \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2 \log(G/\epsilon)}\right) + \frac{\log(1/\delta)}{\epsilon^2}\right). \quad \text{(B1)}$$

Meanwhile, there exists such an algorithm using

$$N = \mathcal{O}\left(\min\left(\frac{2^n \log(1/\delta)}{\epsilon^2}, \frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)\right). \quad \text{(B2)}$$

Here, the minimum with $2^n/\epsilon^2$ corresponds to the sample-optimal approaches for full quantum state tomography [21,22]. The theorem in the main text corresponds to $\delta = \mathcal{O}(1)$, so that the upper and lower bounds are equal up to logarithmic factors.

In Appendix B 1, we prove the sample-complexity upper bound and in Appendix B 2, we show the sample-complexity lower bound. Moreover, in Appendix B 3, we prove Theorem 2, which gives a lower bound on the computational complexity required for this task.

1. Sample-complexity upper bound

In this section, we prove the sample-complexity upper bound for Theorem 12. We provide an algorithm for learning the unknown quantum state within trace distance ϵ by constructing a covering net over the space of all unitaries consisting of G two-qubit gates. We can then obtain a covering net over all pure quantum states generated by G two-qubit gates by applying each element of the unitary covering net to the zero state. With this covering net, we can use quantum hypothesis selection [160] based on classical shadows [110] (discussed in Appendix A 3) to identify a state in the covering net that is close to the unknown target state. We note that this strategy may be adapted to other restricted state or unitary classes as long as we can construct a covering net with bounded cardinality.

Proposition 4 (State-learning upper bound). Let $\epsilon, \delta > 0$. Suppose that we are given N copies of a pure n-qubit state density matrix $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = U|0\rangle^{\otimes n}$ is generated by a unitary U consisting of G two-qubit gates. Then, there exists an algorithm that can output $\hat{\rho}$ such that $d_{\mathrm{tr}}(\hat{\rho}, \rho) \leq \epsilon$ with probability at least $1 - \delta$ using

$$N = \mathcal{O}\left(\min\left(\frac{2^n \log(1/\delta)}{\epsilon^2}, \frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)\right) \quad \text{(B3)}$$

samples of $|\psi\rangle$.

Here, we take the minimum with $2^n/\epsilon^2$, as this is the upper bound achieved for full quantum state tomography on an arbitrary n-qubit pure state [21,22]. Thus, we focus on proving the second term in the minimum. We prove this upper bound by considering two cases: (1) $G \geq n/2$ and (2) $G < n/2$. The upper bounds for each case agree and are given by Eq. (B3). We first prove the proposition for case (1) and indicate what changes for case (2).

Proof of case (1). As previously described, this follows by first creating a covering net over all unitaries consisting of G two-qubit gates and then using quantum hypothesis selection [160].

By Theorem 8, we know that there exists an $(\epsilon/6)$-covering net $\mathcal{N}_{\epsilon/6}$ of the space of unitaries implemented by G two-qubit gates with respect to the diamond distance $d_\diamond$ with metric entropy bounded by

$$\log\left(|\mathcal{N}_{\epsilon/6}|\right) \leq 32G\log\left(\frac{72G}{\epsilon}\right) + 2G\log(n). \tag{B4}$$

Applying each unitary $V' \in \mathcal{N}_{\epsilon/6}$ to the zero state, we obtain a new covering net,

$$\mathcal{N}'_{\epsilon/6} = \{V'|0\rangle\langle 0|^{\otimes n} V'^{\dagger} : V' \in \mathcal{N}_{\epsilon/6}\}, \tag{B5}$$

for the set of pure quantum states generated by G two-qubit gates with respect to the trace distance. We argue that this is true as follows. Any pure quantum state generated by G two-qubit gates can be written as $|\phi\rangle = V|0\rangle^{\otimes n}$ for some unitary V implemented by G two-qubit gates and we let $\sigma = |\phi\rangle\langle\phi|$. Using the definition of the covering net $\mathcal{N}_{\epsilon/6}$, there exists a unitary $V' \in \mathcal{N}_{\epsilon/6}$ such that $d_\diamond(V, V') < \epsilon/6$. Consider $|\phi'\rangle = V'|0\rangle^{\otimes n}$ and let $\sigma' = |\phi'\rangle\langle\phi'| \in \mathcal{N}'_{\epsilon/6}$. By the definition of the diamond distance in terms of a worst case over input states, we also have $d_{\mathrm{tr}}(\sigma, \sigma') \leq d_\diamond(V, V') \leq \epsilon/12 < \epsilon/6$. Thus, $\mathcal{N}'_{\epsilon/6}$ satisfies the definition of a covering net over the pure quantum states generated by G two-qubit gates with respect to the trace distance d_{tr}. Moreover, we clearly see that $|\mathcal{N}'_{\epsilon/6}| \leq |\mathcal{N}_{\epsilon/6}|$.

We can consider this covering net $\mathcal{N}'_{\epsilon/6}$ as the set of hypothesis states in Proposition 1. Let $\rho = |\psi\rangle\langle\psi|$ be the unknown quantum state of which we have copies. By Proposition 1, there exists an algorithm to learn $\tilde{\rho}$ such that

$$d_{\mathrm{tr}}(\rho, \tilde{\rho}) \leq 3 \cdot \frac{\epsilon}{6} + \frac{\epsilon}{2} = \epsilon \tag{B6}$$

with probability at least $1 - \delta$. Here, note that we have used $\eta = \epsilon/6$ in Proposition 1 by the definition of an $(\epsilon/6)$-covering net. Furthermore, we may choose $\epsilon_2 = \epsilon/2$ and $\delta_1 = \delta/2$. In this way, we obtain $\tilde{\rho}$ such that $d_{\mathrm{tr}}(\rho, \tilde{\rho}) \leq \epsilon$ with probability at least $1 - \delta$. Moreover, by Proposition 1, this algorithm to find $\tilde{\rho}$ requires at most

$$\begin{aligned} N &= \mathcal{O}\left(\frac{\log\left(|\mathcal{N}'_{\epsilon/6}|/\delta\right)}{\epsilon^2}\right) \\ &= \mathcal{O}\left(\frac{G\log(G/\epsilon) + G\log(n) + \log(1/\delta)}{\epsilon^2}\right) \end{aligned} \tag{B7}$$

copies of ρ, where the second equality follows from Eq. (B4). Because we are considering $G \geq n/2$ in this case, we then have

$$N = \mathcal{O}\left(\frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^2}\right), \tag{B8}$$

as claimed. ■

Note that in the above proof we have used $G \geq n/2$ in the last step to remove the extra $\log(n)$ factor. However, in case (2), we can no longer execute this step and we must consider a more careful strategy to remove the dependence on system size n. The key observation is that if $G < n/2$, some qubits in the system will be left in the zero state because no gate has acted upon them (for G two-qubit gates, at most $2G < n$ qubits are acted upon nontrivially). Note that we only need to learn the quantum state on these $2G$ qubits rather than the whole system, since we can simply tensor product with the zero state for the remaining qubits. Thus, we require the ability to discern which qubits have been acted upon by the G two-qubit gates. Once we find this set of qubits, the idea is to consider a covering net for the set of pure quantum states generated by G two-qubit gates *on this restricted system*. Then, we can follow a similar argument to the above proof of case (1).

We prove case (2) of Proposition 4 in the following sections. For the rest of this section, let $\rho = |\psi\rangle\langle\psi|$. In Appendix B 1 a, we discuss an algorithm that identifies the qubits acted on nontrivially by the G two-qubit gates with high probability and show that restricting to these identified qubits does not cause much error. In Appendix B 1 b, we resolve a technical issue for defining the covering net on the restricted system, which stems from the algorithm possibly not identifying all qubits. Finally, in Appendix B 1 c, we combine these pieces to provide the full proof of case (2).

a. Postselection

First, we present an algorithm to determine which qubits of the unknown quantum state $\rho = |\psi\rangle\langle\psi|$ have been acted upon nontrivially by the G two-qubit gates. We then prove a guarantee about the number of samples of ρ needed to determine these qubits with high probability. We also show that considering ρ to be the zero state on the rest of the qubits does not incur much error.

Suppose that the true set of qubits acted upon by the G two-qubit gates is denoted as A. To determine which

ALGORITHM 1. Identify qubits acted upon nontrivially (state version).

Input: Copies of unknown n-qubit quantum state ρ.
Output: List $\hat{A} \subseteq [n]$ of qubits.
1 Initialize $\hat{A} = \emptyset$.
2 Repeat the following $N = \mathcal{O}\left(\frac{G+\log(1/\delta_1)}{\epsilon_1}\right)$ times:
(a) Measure all qubits of the unknown state ρ in the computational basis.
(b) Given the measurement outcome $|x\rangle$, set $\hat{A} \leftarrow \hat{A} \cup \mathrm{supp}(x)$, where $\mathrm{supp}(x) = \{i \in [n] : x_i \neq 0\}$.

qubits are in the set A, consider the procedure given in Algorithm 1.

The idea behind this algorithm is simple. If we measure a qubit in the computational basis and receive a nonzero measurement outcome, then it must have been acted upon by one of the G two-qubit gates, because the quantum state is assumed to have been initialized in the zero state. We prove that $\mathcal{O}(G + \log(1/\delta_1)/\epsilon_1)$ copies of ρ suffice to obtain, with high probability $1 - \delta_1$, the desired property that measuring the qubits in $\hat{B} \triangleq [n] \setminus \hat{A}$ of ρ yields the all-zero bit string with high probability $1 - \epsilon_1$.

Lemma 11. Let $\epsilon_1, \delta_1 > 0$. Suppose that we are given copies of a pure n-qubit quantum state $\rho = |\psi\rangle\langle\psi|$ generated by G two-qubit gates acting on a subset of the qubits $A \subseteq [n]$. Then, Algorithm 1 uses $N = \mathcal{O}(G + \log(1/\delta_1)/\epsilon_1)$ copies of ρ and outputs with probability at least $1 - \delta_1$ a list $\hat{A} \subset [n]$ such that

$$\left\langle 0_{\hat{B}} \right| \rho_{\hat{B}} \left| 0_{\hat{B}} \right\rangle \geq 1 - \epsilon_1, \tag{B9}$$

where $\rho_{\hat{B}}$ denotes the reduced density matrix of ρ when tracing out all qubits other than those in the set $\hat{B} = [n] \setminus \hat{A}$ and $\left|0_{\hat{B}}\right\rangle$ denotes the zero state on all qubits in $\hat{B}$.

Proof. Let A' be any possible set that could be output by Algorithm 1. Let $B' \triangleq [n] \setminus A'$. We first define some random variables to state our claim more precisely. Let $E_{i,A'}$ be the event that round i of measurement of the qubits in $B' = [n] \setminus A'$ in Algorithm 1 yields the all-zero bit string. Let $X_{i,A'}$ be the indicator random variable corresponding to the event $E_{i,A'}$. Then, we have that $\bar{X}_{A'} \triangleq 1/N \sum_{i=1}^{N} X_{i,A'}$ is the number of times the qubits in B' are all measured to be zero divided by the total number of measurements. In other words, $\bar{X}_{A'}$ is an empirical estimate for the overlap that the state $\rho_{B'}$ on qubits in B' has with the all-zero state. Moreover, we have

$$\mathbb{E}[X_{A'}] \triangleq \mathbb{E}[X_{i,A'}] = \langle 0_{B'} | \rho_{B'} | 0_{B'} \rangle, \tag{B10}$$

for all A'. Note that the first definition makes sense because, for any i, the $X_{i,A'}$ are identically distributed. This says that the true expectation of our random variables is the true overlap of the state $\rho_{B'}$ with the all-zero state.

We claim that for any A', if the true overlap is less than $1 - \epsilon_1$, then the estimated overlap is less than $1 - \epsilon_1/2$ with high probability. Formally, in terms of our random variables, this is the following statement.

Claim 1. For any set A' that could be output by Algorithm 1, if $\mathbb{E}[X_{A'}] < 1 - \epsilon_1$, then $\bar{X}_{A'} < 1 - \epsilon_1/2$ with probability at least $1 - \delta_1$.

Thus, we have reduced our task to a concentration problem. Note that it suffices to prove this because the set $\hat{A}$ actually identified by Algorithm 1 has $\bar{X}_{\hat{A}} = 1$. This is true because a qubit is only added to the set $\hat{A}$ in the algorithm if it has measured and observed a nonzero outcome. Thus, all qubits in $\hat{B} = [n] \setminus \hat{A}$ must have given zero when measured throughout all rounds of measurement. By definition, this gives us that $\bar{X}_{\hat{A}} = 1$. Then, by the contrapositive of Claim 1, we see that $\mathbb{E}[X_{\hat{A}}] = \left\langle 0_{\hat{B}} \right| \rho_{\hat{B}} \left| 0_{\hat{B}} \right\rangle \geq 1 - \epsilon_1$ with probability at least $1 - \delta_1$. We now prove this claim using classical concentration inequalities. ■

Proof of Claim 1. First, we fix some set A' that could be output by Algorithm 1. Suppose that

$$\mathbb{E}[X_{A'}] \triangleq 1 - a < 1 - \epsilon_1, \tag{B11}$$

where $a > \epsilon_1$. Recall the Bhatia-Davis inequality, which states that, for $X \in [b, d]$,

$$\mathrm{Var}(X) \leq (d - \mathbb{E}[X])(\mathbb{E}[X] - b). \tag{B12}$$

In our case, we have $X_{A'} \in [0, 1]$, since they are indicator random variables, so that the inequality gives us

$$\mathrm{Var}(X_{A'}) \leq (1 - \mathbb{E}[X])\mathbb{E}[X] \leq 1 - \mathbb{E}[X] = a. \tag{B13}$$

Now, recall Bernstein's inequality, which states that for independent random variables X_i with $|X_i| \leq c$ and

$\sigma^2 = 1/N \sum_{i=1}^{N} \mathrm{Var}(X_i)$, we have, for any $t > 0$,

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X] > t\right) \leq \exp\left(-\frac{Nt^2}{2\sigma^2 + 2ct/3}\right). \tag{B14}$$

In our case, $c = 1, \sigma^2 \leq a$, and $t = a/2$. Then, Bernstein's inequality results in

$$\Pr\left(\bar{X}_{A'} - \mathbb{E}[X] > \frac{a}{2}\right) \leq \exp\left(-\frac{Na^2/4}{2a + a/3}\right). \tag{B15}$$

Plugging in $\mathbb{E}[X] = 1 - a$ and simplifying, we have

$$\Pr\left(\bar{X}_{A'} > 1 - \frac{a}{2}\right) \leq \exp\left(-\frac{3Na}{28}\right) \leq \exp\left(-\frac{3N\epsilon_1}{28}\right). \tag{B16}$$

Since $a > \epsilon_1$, then $1 - a/2 < 1 - \epsilon_1/2$, so that we have

$$\Pr\left(\bar{X}_{A'} > 1 - \frac{\epsilon_1}{2}\right) \leq \exp\left(-\frac{3N\epsilon_1}{28}\right). \tag{B17}$$

Plugging in $N = 28\log\left(2^{2G}/\delta_1\right)/3\epsilon_1$, we have

$$\Pr\left(\bar{X}_{A'} > 1 - \frac{\epsilon_1}{2}\right) \leq \frac{\delta_1}{2^{2G}}. \tag{B18}$$

Recall that this inequality was for a single fixed set A' but we want our claim to hold for any set A'. Thus, we need to apply the union bound over all possible sets A' output by Algorithm 1.

We claim that the number of such sets is at most 2^{2G}. This is clear because if A' is output by the algorithm, then $A' \subseteq A$, where A is the true set of qubits on which the G gates act nontrivially. This is true by construction because in order for a qubit to be added to the set output by Algorithm 1, its result upon measurement must have yielded a nonzero outcome, so that a gate must have acted upon this qubit. Hence $A' \subseteq A$, and because $|A| \leq 2G$, the number of possible subsets A' of A is at most 2^{2G}.

Thus, applying a union bound to Eq. (B18), we see that the probability that, for any A', $\bar{X}_{A'}$ is greater than $1 - \epsilon_1/2$ is at most δ_1. In other words, $\bar{X}_{A'}$ is less than $1 - \epsilon_1/2$ with probability at least $1 - \delta_1$. Moreover, here we have used

$$N = \frac{28\log\left(2^{2G}/\delta_1\right)}{3\epsilon_1} = \mathcal{O}\left(\frac{G + \log(1/\delta_1)}{\epsilon_1}\right). \tag{B19}$$

This concludes the proof of the claim, which gives the result in Lemma 11 as explained previously. ■

With this, we know that measuring qubits in $\hat{B} = [n] \setminus \hat{A}$ of ρ yields the all-zero bit string with high probability. We want to show that, in fact, we can consider $\rho_{\hat{B}}$ as being the zero state without incurring much error. In particular, we want to show the following lemma.

Lemma 12. Let $\epsilon, \delta_1 > 0$. Suppose that we are given $N = \mathcal{O}\left(G + \log(1/\delta_1)/\epsilon^2\right)$ copies of an n-qubit quantum state ρ generated by G gates. Let $\hat{A} \subset [n]$ be as in Algorithm 1 and let $\hat{B} = [n] \setminus \hat{A}$. Then, for $\Lambda = \left|0_{\hat{B}}\right\rangle\left\langle 0_{\hat{B}}\right| \otimes I_{\hat{A}}$ (where $\left|0_{\hat{B}}\right\rangle$ denotes the zero state on all qubits in $\hat{B}$) and for the postmeasurement state

$$\rho' \triangleq \frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\rho)}, \tag{B20}$$

we have

$$d_{\mathrm{tr}}(\rho, \rho') \leq \frac{\epsilon}{24} \tag{B21}$$

with probability at least $1 - \delta_1$.

In other words, we want to show that our original state ρ is not far in trace distance from the new state ρ', where ρ' is the state ρ with the qubits in $\hat{B}$ projected to the zero state. In this way, we can effectively only consider the system on qubits in $\hat{A}$ when defining the covering net and using hypothesis selection. This turns out to be a bit more nuanced, but this is the general idea. To show this, we will use the "gentle measurement lemma," following the presentation in Ref. [155].

Lemma 13 (Lemma 9.4.1 in Ref. [155]). Consider a density operator ρ and a measurement operator Λ, where $0 \leq \Lambda \leq I$. The measurement operator could be an element of a positive operator-valued measure (POVM). Suppose that the measurement operator Λ has a high probability of detecting the state ρ:

$$\mathrm{Tr}(\Lambda\rho) \geq 1 - \epsilon, \tag{B22}$$

where $\epsilon \in [0, 1]$ (the probability of detection is high if ϵ is close to zero). Then, the postmeasurement state

$$\rho' \triangleq \frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\rho)} \tag{B23}$$

is $\sqrt{\epsilon}$-close to the original state ρ in trace distance:

$$d_{\mathrm{tr}}(\rho, \rho') \leq \sqrt{\epsilon}. \tag{B24}$$

Thus, the measurement does not disturb the state ρ by much if ϵ is small.

With this, we can now prove Lemma 12.

Proof of Lemma 12. As stated above, let $\hat{A} \subset [n]$ be as in Algorithm 1 and let $A \subset [n]$ be the true set of qubits acted nontrivially on by the G gates. Let $\hat{B} \triangleq [n] \setminus \hat{A}$ and let $B \triangleq [n] \setminus A$.

In order to apply the gentle measurement lemma, we need to show that

$$\mathrm{Tr}(\Lambda\rho) \geq 1 - \left(\frac{\epsilon}{24}\right)^2. \tag{B25}$$

Since $\Lambda = |0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}}$, where $|0_{\hat{B}}\rangle$ denotes the zero state on all qubits in $\hat{B}$, we have

$$\begin{aligned}\mathrm{Tr}(\Lambda\rho) &= \mathrm{Tr}((|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}})\rho) \\ &= \mathrm{Tr}(|0_{\hat{B}}\rangle\langle 0_{\hat{B}}|\,\rho_{\hat{B}}) = \langle 0_{\hat{B}}|\,\rho_{\hat{B}}\,|0_{\hat{B}}\rangle,\end{aligned} \tag{B26}$$

where $\rho_{\hat{B}}$ denotes the reduced density matrix obtained by tracing out all qubits in $[n] \setminus \hat{B}$. Thus, it suffices to show that

$$\langle 0_{\hat{B}}|\,\rho_{\hat{B}}\,|0_{\hat{B}}\rangle \geq 1 - \left(\frac{\epsilon}{24}\right)^2. \tag{B27}$$

Intuitively, this makes sense because in Algorithm 1, we have identified the qubits in $\hat{B}$ as those being close to the zero state. Indeed, this holds by Lemma 11 when choosing $\epsilon_1 = (\epsilon/24)^2$. Thus, the result follows. ■

b. Permutation

Before we can prove Proposition 4, we must resolve a technical issue. Namely, we would ideally like to consider a covering net on the subsystem of qubits in the set A (the true set of qubits on which the G gates generating the unknown state ρ act nontrivially). In this way, because $\hat{A} \subseteq A$, where $\hat{A}$ is the set of qubits identified by Algorithm 1, then our postselected state ρ' from Lemma 12 should be close to some state in this covering net on the subsystem. This nearby state in the covering net can then be identified via quantum hypothesis selection [54]. By Lemma 12, this state from hypothesis selection is also close to the original unknown state ρ.

However, the problem with the above is that we do not know the true set of qubits A; we only know the identified set of qubits $\hat{A}$. Moreover, it is possible that $\hat{A} \subsetneq A$, i.e., Algorithm 1 may not have been able to detect certain qubits as having been acted upon nontrivially by the G gates. For example, suppose that when preparing the unknown state ρ, certain qubits are used as workspace ancillas and are reset to the zero state at the end of the computation.

In order to define a covering net on a system on which the G gates act (the setting of Lemma 9), we need to somehow identify the qubits in $A \setminus \hat{A}$ that are undetected by the algorithm. To do so, we argue that we can permute the qubits outside of the set $\hat{A}$ and not deviate much from the original state ρ. In this way, without loss of generality, we can permute the qubits such that those in $A \setminus \hat{A}$ are grouped together in some fixed set of qubits. Then, we can define a covering net on the system of qubits defined by this fixed set containing the qubits in $A \setminus \hat{A}$ and our identified set $\hat{A}$. By construction, we know that the G gates act on this subset of qubits, so this is the correct setting of Lemma 9. We note that the permutations used in the proof are a mathematical tool for the analysis but the learner has to neither know nor perform these permutations.

To formalize this, we first define a permutation and claim that permuting the qubits outside of the set $\hat{A}$ does not change the postselected state ρ'.

Definition 11 (Permutation). A unitary $W \in \mathrm{U}(2^n)$ is a *permutation unitary* if it satisfies the following property: W corresponds to a permutation $\sigma_W \in S_n$ of order 2, where S_n is the symmetric group of size n, and W acts as

$$W|x_1 \ldots x_n\rangle = \left|x_{\sigma_W(1)} \cdots x_{\sigma_W(n)}\right\rangle, \tag{B28}$$

where $x = x_1 \cdots x_n \in \{0,1\}^n$. Moreover, we use W_S for a set $S \subseteq \{1, \ldots, n\}$ to denote a permutation unitary where the corresponding permutation σ_{W_S} is such that $\sigma_{W_S}|_{\overline{S}} = \mathrm{id}$, where $\overline{S} = [n] \setminus S$. In other words, σ_{W_S} only permutes the elements in S.

It is easy to see here that because the corresponding permutation is of order 2, W is Hermitian. Our next lemma shows that such permutations, when acting only on $\hat{B}$, do not change our postselected state.

Lemma 14. Let ρ' be as in Lemma 12. Explicitly, let $\hat{A} \subset [n]$ be as in Algorithm 1 and let $\hat{B} = [n] \setminus \hat{A}$. Then, for $\Lambda = |0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}}$ (where $|0_{\hat{B}}\rangle$ denotes the zero state on all qubits in $\hat{B}$), define

$$\rho' = \frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\rho)}. \tag{B29}$$

Then, we have

$$\rho'' \triangleq W_{\hat{B}}\rho' W_{\hat{B}} = \rho', \tag{B30}$$

where $W_{\hat{B}}$ is any permutation unitary that only permutes qubits in $\hat{B}$.

Proof. To see the claim, we can simply expand the expression for ρ'':

$$\rho'' = W_{\hat{B}}\rho' W_{\hat{B}} \tag{B31}$$

$$= W_{\hat{B}}\frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\rho)}W_{\hat{B}} \tag{B32}$$

$$= W_{\hat{B}}\frac{\Lambda\rho\Lambda}{\mathrm{Tr}(\Lambda\rho)}W_{\hat{B}} \tag{B33}$$

$$= \frac{W_{\hat{B}}(|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I)\rho(|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I)W_{\hat{B}}}{\mathrm{Tr}(\Lambda\rho)} \tag{B34}$$

$$= \frac{(|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I)\rho(|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I)}{\mathrm{Tr}(\Lambda\rho)} \tag{B35}$$

$$= \rho', \tag{B36}$$

where in the third line we have used that Λ is a projector, so that $\sqrt{\Lambda} = \Lambda$, and in the fifth line, we have used that $W_{\hat{B}}$ only permutes the qubits in $\hat{B}$, which does not have any effect because, here, all qubits in $\hat{B}$ are in the zero state. ■

Lemma 15. Let $\epsilon, \delta_2 > 0$. The trace distance between ρ and the permuted state $\tilde{\rho} = W_{\hat{B}}\rho W_{\hat{B}}$, where $W_{\hat{B}}$ is any permutation unitary that only permutes qubits in $\hat{B}$, is less than $\epsilon/24$:

$$d_{\mathrm{tr}}(\rho, \tilde{\rho}) \leq \frac{\epsilon}{12} \tag{B37}$$

with probability at least $1 - \delta_2$.

Proof. This proof combines Lemmas 14 and 12. The idea is the following. We know from Lemma 12 that ρ and the postselected state ρ' are close in trace distance. Moreover, by Lemma 14, we know that the postselected state ρ' and the permuted postselected state ρ'' are equal (without error). We can also show, similarly to Lemma 12, that the permuted state $\tilde{\rho}$ is close to the postselected state $\tilde{\rho}'$, where this postselection is done in the same way as Lemma 12 by replacing ρ with $\tilde{\rho}$. Moreover, we can see that $\rho'' = \tilde{\rho}'$, so the claim then follows by the triangle inequality.

Now, let us formalize this. By Lemma 12, we have

$$d_{\mathrm{tr}}(\rho, \rho') \leq \frac{\epsilon}{24} \tag{B38}$$

with probability at least $1 - \delta_2/2$ (choosing $\delta_1 = \delta_2/2$), where

$$\rho' = \frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\rho)} \tag{B39}$$

for $\Lambda = |0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}}$. By Lemma 14, we know that

$$\rho'' \triangleq W_{\hat{B}}\rho' W_{\hat{B}} = \rho', \tag{B40}$$

where $W_{\hat{B}}$ is a permutation that only affects qubits in $\hat{B}$. Now, consider the permuted state $\tilde{\rho} = W_{\hat{B}}\rho W_{\hat{B}}$. Recall that in the proof of Lemma 12, to obtain Eq. (B38), it sufficed to show that $\mathrm{Tr}(\Lambda\rho) \geq 1 - (\epsilon/24)^2$ and the result followed by the gentle measurement lemma (Lemma 13). Thus, by the same proof, as long as $\mathrm{Tr}(\Lambda\tilde{\rho}) \geq 1 - (\epsilon/24)^2$, then we also have

$$d_{\mathrm{tr}}(\tilde{\rho}, \tilde{\rho}') \leq \frac{\epsilon}{24} \tag{B41}$$

with probability at least $1 - \delta_2/2$, where

$$\tilde{\rho}' \triangleq \frac{\sqrt{\Lambda}\tilde{\rho}\sqrt{\Lambda}}{\mathrm{Tr}(\Lambda\tilde{\rho})}. \tag{B42}$$

We can clearly see that this condition holds:

$$\begin{aligned}\mathrm{Tr}(\Lambda\tilde{\rho}) &= \mathrm{Tr}((|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}})W_{\hat{B}}\rho W_{\hat{B}}) \\ &= \mathrm{Tr}((|0_{\hat{B}}\rangle\langle 0_{\hat{B}}| \otimes I_{\hat{A}})\rho) = \mathrm{Tr}(\Lambda\rho) \geq 1 - (\epsilon/24)^2,\end{aligned} \tag{B43}$$

where the second equality follows because $W_{\hat{B}}$ only permutes qubits in $\hat{B}$, which (rearranging with the trace) does not have any effect on $|0_{\hat{B}}\rangle\langle 0_{\hat{B}}|$ because all qubits in $\hat{B}$ are in the zero state. Thus, Eq. (B41) holds.

We also claim that $\rho'' = \tilde{\rho}'$. This follows by effectively the same proof as Lemma 14.

Putting everything together, we have that $\rho' = \rho'' = \tilde{\rho}'$. Thus, by Eq. (B38),

$$d_{\mathrm{tr}}(\rho, \tilde{\rho}') \leq \frac{\epsilon}{24} \tag{B44}$$

with probability at least $1 - \delta_2/2$. By the triangle inequality with Eq. (B41), we then obtain the claim:

$$d_{\mathrm{tr}}(\rho, \tilde{\rho}) \leq \frac{\epsilon}{12} \tag{B45}$$

with probability at least $1 - \delta_2$. ■

c. Proof of case (2) of Proposition 4

With this, we can prove case (2) of Proposition 4. Recall that in case (2), we require that $G < n/2$. We have provided a sketch of the argument throughout the previous sections, so we put everything together here.

Proof of case (2) of Proposition 4. Let $\epsilon, \delta > 0$. Consider $G < n/2$. Because G is small compared to n, there exist some qubits that have not been acted upon by the G gates used to generate the state $\rho = |\psi\rangle\langle\psi|$. Thus, since we assume that the unknown quantum state ρ is constructed by applying a unitary to the all-zero state, then these qubits not acted upon by the G gates remain in the zero state. Using the techniques in Appendix B 1 a, we can find the qubits that are acted on nontrivially by the G gates. Then, we want to consider the covering net on only this set of qubits. However, because our algorithm does not necessarily find *all* qubits acted on nontrivially by the G

gates, we argue in Appendix B 1 b that we can permute the qubits in the system without significantly affecting the original state ρ. In this way, we can consider a permutation that gathers those qubits acted upon nontrivially that our algorithm did not find into some fixed set. We can then define the covering net on the subsystem consisting of this fixed set along with the identified set of qubits.

Let us now formalize these ideas. Let $\hat{A}$ be the set of qubits identified by Algorithm 1 and let A be the true set of qubits acted on nontrivially by the G gates. Let $W_{\hat{B}}$ be a permutation only affecting the qubits in $\hat{B} \triangleq [n] \setminus A$ (Definition 11), which gathers the qubits in $A \setminus \hat{A}$ into some fixed set of qubits C. Since $|C| + |\hat{A}| = |A \setminus \hat{A}| + |\hat{A}| = |A| \leq 2G$, then $C \cup \hat{A}$ has at most $2G$ qubits and these qubits are acted upon by G gates.

By Theorem 8, we know that there exists an $(\epsilon/12)$-covering net $\mathcal{N}_{\epsilon/12}$ of the space of unitaries implemented by G two-qubit gates on the permuted system consisting of only qubits in $C \cup \hat{A}$ with respect to the diamond distance $d_\diamond = \max_\rho \|(U \otimes I)\rho(U \otimes I)^\dagger - (V \otimes I)\rho(V \otimes I)^\dagger\|_1$. Moreover, this covering net has metric entropy bounded by

$$\log\left(|\mathcal{N}_{\epsilon/12}|\right) \leq 32G \log\left(\frac{144G}{\epsilon}\right) + 2G\log(2G) = \mathcal{O}\left(G\log(G/\epsilon)\right). \tag{B46}$$

We can instead consider

$$\mathcal{N}'_{\epsilon/12} = \{V' \left|0_{C\cup\hat{A}}\right\rangle\left\langle 0_{C\cup\hat{A}}\right| V'^\dagger : V' \in \mathcal{N}_{\epsilon/12}\}, \tag{B47}$$

where $\left|0_{C\cup\hat{A}}\right\rangle$ denotes the zero state on all qubits in our subsystem $C \cup \hat{A}$. By the same argument as in case (1), $\mathcal{N}'_{\epsilon/12}$ defines a covering net over the set of pure quantum states on the subsystem $C \cup \hat{A}$ generated by G two-qubit gates with respect to the trace distance. Moreover, $|\mathcal{N}'_{\epsilon/12}| \leq |\mathcal{N}_{\epsilon/12}|$.

Since this covering net $\mathcal{N}'_{\epsilon/12}$ is only for states on at most $2G$ qubits, let $\mathcal{N}''_{\epsilon/12}$ be the set of states where each state in $\mathcal{N}'_{\epsilon/12}$ is tensored with the zero state for qubits in $[n] \setminus (C \cup \hat{A})$. Let $\tilde{\rho} = W_{\hat{B}}\rho W_{\hat{B}}$ be the original state on this permuted system. By the definition of a covering net, we know that there exists some $\sigma_i \in \mathcal{N}''_{\epsilon/12}$ such that

$$d_{\mathrm{tr}}(\tilde{\rho}, \sigma_i) \leq \frac{\epsilon}{12}. \tag{B48}$$

We justify this further in the following. By definition, the only qubits in the state $\tilde{\rho}$ that are acted on nontrivially by the G gates are those in $C \cup \hat{A}$. Since no gates act on qubits outside of $C \cup \hat{A}$, then the other qubits in $\tilde{\rho}$ must be in the zero state. Hence, we can write $\tilde{\rho} = \tilde{\rho}_{C\cup\hat{A}} \otimes |0\rangle\langle 0|^{\otimes(n-|C\cup\hat{A}|)}$, where $\tilde{\rho}_{C\cup\hat{A}}$ denotes the state of the qubits in $C \cup \hat{A}$ that are acted upon by the G gates. Moreover, by the definition of a covering net, then there exists some $\sigma_{i,C\cup\hat{A}} \in \mathcal{N}'_{\epsilon/12}$ such that

$$d_{\mathrm{tr}}(\tilde{\rho}_{C\cup\hat{A}}, \sigma_{i,C\cup\hat{A}}) \leq \frac{\epsilon}{12}, \tag{B49}$$

where, similarly, $\sigma_{i,C\cup\hat{A}}$ is a state on the qubits in $C \cup \hat{A}$ that are acted upon by G gates. Taking the tensor product with the zero state on the remaining qubits does not affect the trace distance. Thus, we can write $\sigma_i = \sigma_{i,C\cup\hat{A}} \otimes |0\rangle\langle 0|^{\otimes(n-|C\cup\hat{A}|)} \in \mathcal{N}''_{\epsilon/12}$, where this satisfies

$$d_{\mathrm{tr}}(\tilde{\rho}, \sigma_i) = d_{\mathrm{tr}}(\tilde{\rho}_{C\cup\hat{A}}, \sigma_{i,C\cup\hat{A}}) \leq \frac{\epsilon}{12}, \tag{B50}$$

as claimed. Moreover, by Lemma 15, choosing $\delta_2 = \delta/2$, we know that

$$d_{\mathrm{tr}}(\rho, \tilde{\rho}) \leq \frac{\epsilon}{12} \tag{B51}$$

with probability at least $1 - \delta/2$. Recall that this approximation requires only

$$N_1 = \mathcal{O}\left(\frac{G + \log(1/\delta)}{\epsilon^2}\right) \tag{B52}$$

copies of ρ (from Lemma 12) for identifying the set $\hat{A}$. By the triangle inequality, we have that there exists some $\sigma_i \in \mathcal{N}''_{\epsilon/12}$ such that

$$d_{\mathrm{tr}}(\rho, \sigma_i) \leq \frac{\epsilon}{6} \tag{B53}$$

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

Using hypothesis selection on the covering net $\mathcal{N}''_{\epsilon/12}$ and the unknown state ρ, by Proposition 1, there exists an algorithm to learn σ such that

$$d_{\mathrm{tr}}(\rho, \sigma) \leq \epsilon \tag{B54}$$

with probability at least $1 - \delta$, where we have chosen $\eta = \epsilon/6$ and $\epsilon/2, \delta/2$ for the parameters in Proposition 1. Moreover, by Proposition 1 and Eq. (B46), this algorithm requires only

$$N_2 = \mathcal{O}\left(\frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^2}\right) \tag{B55}$$

copies of ρ. Putting everything together, we have that

$$d_{\mathrm{tr}}(\rho, \sigma) \leq \epsilon \tag{B56}$$

with probability at least $1-\delta$, where our algorithm to find σ requires only

$$N = N_1 + N_2 = \mathcal{O}\left(\frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^2}\right). \quad \text{(B57)}$$

This matches our upper bound for case (1) and thus concludes the proof of Proposition 4. ■

2. Sample-complexity lower bound

In this section, we prove the sample-complexity lower bound for Theorem 12.

Proposition 5 (State-learning lower bound). Let $\epsilon, \delta > 0$. Suppose that we are given N copies of an n-qubit pure-state density matrix $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = U|0\rangle^{\otimes n}$ is generated by a unitary U consisting of G two-qubit gates. Then, any algorithm that can output $\hat{\rho}$ such that $d_{\mathrm{tr}}(\hat{\rho}, \rho) \leq \epsilon$ with probability at least $1-\delta$ requires at least

$$N = \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2\log(G/\epsilon)}\right) + \frac{\log(1/\delta)}{\epsilon^2}\right) \quad \text{(B58)}$$

samples of $|\psi\rangle$.

Here, similarly to the upper bound, we take the minimum with $\Omega(2^n/\epsilon^2)$, as this is the lower bound achieved for full quantum state tomography [21,22]. We thus focus on the second term in the minimum. We first consider the number of samples required to learn n-qubit pure quantum states generated by G gates *applied only to the first* $\lfloor \log_2(G/C) \rfloor$ *qubits* (for some constant C specified later) of the $n \geq \lfloor \log_2(G/C) \rfloor$ qubits in total. Denote this set of states as S_1. Note that if $n \leq \lfloor \log_2(G/C) \rfloor$, then we can simply import the lower bound for full quantum state tomography [21,22]. We later reduce the general case, in which the G gates can be applied on any of the qubits, to this case. Namely, we prove the following proposition.

Proposition 6. Let $\epsilon, \delta > 0$. Suppose that we are given N copies of an n-qubit pure-state density matrix $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = (U \otimes I)|0\rangle^{\otimes n} \in S_1$ is generated by a unitary U consisting of G two-qubit gates applied only to the first $\lfloor \log_2(G/C) \rfloor$ qubits for some constant C. Then, any algorithm that can output $\hat{\rho}$ such that $d_{\mathrm{tr}}(\hat{\rho}, \rho) \leq \epsilon$ with probability at least $1-\delta$ requires at least

$$N = \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2\log(G/\epsilon)}\right) + \frac{\log(1/\delta)}{\epsilon^2}\right) \quad \text{(B59)}$$

samples of $|\psi\rangle$.

We note that for constant error ϵ, the $\Omega(G/\log G)$ lower bound can be improved to $\Omega(G)$ using Refs. [173,174].

We prove Proposition 6 by combining results from Refs. [21,143]. Namely, the lower bound in Ref. [21] works by lower bounding the sample complexity of learning any rank r d-dimensional quantum state in terms of the packing number of this space of states. We apply the authors' results to our setting, where the space of states that the packing net is defined over is S_1 instead. We first recall important results from Refs. [21,143] that we use throughout the proof. In Ref. [21], the sample complexity of learning a d-dimensional pure state is lower bounded as follows.

Theorem 13 (In proof of Theorem 3 in Ref. [21]). Let $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Suppose that there exists a POVM $\{M_\sigma d\sigma\}$ on $(\mathbb{C}^d)^{\otimes N}$ such that for a pure quantum state $\rho \in \mathbb{C}^{d\times d}$,

$$\int_{d_{\mathrm{tr}}(\sigma,\rho)\leq\epsilon} d\sigma\,\mathrm{Tr}[M_\sigma \rho^{\otimes N}] \geq 1-\delta. \quad \text{(B60)}$$

Then,

$$N \geq \frac{(1-\delta)\ln m - \ln 2}{\chi_0}, \quad \text{(B61)}$$

where m is the size of an (2ϵ)-packing net of the space of d-dimensional pure-state density matrices and

$$\chi_0 \triangleq S(\mathbb{E}_U[U\rho_x U]) - S(\rho_x) \quad \text{(B62)}$$

is the Holevo information, where ρ_x is any element of the (2ϵ)-packing net, S is the von Neumann entropy, and the expectation is taken over the Haar measure.

This states that any measurement procedure that can identify a state ρ up to ϵ trace distance requires at least N copies of ρ, where N is given by Eq. (B61) and depends on the size of an (2ϵ)-packing net of the space of d-dimensional pure-state density matrices. Moreover, in Ref. [21], the size of such a packing net has been bounded.

Lemma 16 (Lemma 5 in Ref. [21]). There exists an ϵ-packing net $\{\rho_1, \ldots, \rho_m\}$ of the space of d-dimensional pure-state density matrices satisfying

$$c\ln m \geq d, \quad \text{(B63)}$$

for c a sufficiently large constant and $d > 3$. This packing net also satisfies

$$\frac{\chi_0}{c} \leq \epsilon^2 \ln\left(\frac{d}{\epsilon}\right) \quad \text{(B64)}$$

for a sufficiently large constant $c > 0$, where χ_0 is given by Eq. (B62).

Finally, the last result we will need gives a bound on the number of gates needed to generate an arbitrary n-qubit pure state.

Lemma 17 (Section 4 of Ref. [143]). Any n-qubit pure quantum state can be recursively defined as the result of a quantum circuit implemented by $\mathcal{O}(2^n)$ two-qubit gates applied to the $|0\rangle^{\otimes n}$ state. Explicitly, this quantum circuit has at most $C \cdot 2^n$ two-qubit gates for some constant C.

With these results, we can prove Proposition 6. The idea is that, using Lemma 17, any pure state on the first $k \sim \log_2 G$ qubits can be generated by G gates. Then, we can use the same packing-net construction as Ref. [21] from Lemma 16. Plugging into Theorem 13 then gives our lower bound. We also add an additional term to account for expected asymptotic δ behavior.

Proof of Proposition 6. We wish to construct a (2ϵ)-packing net over the space S_1 of n-qubit pure quantum states generated by applying G gates to the first $k = \lfloor \log_2(G/C) \rfloor$ qubits, where C is taken to be the same constant as in Lemma 17. First, consider only the subsystem consisting of the first k qubits. Note that by Lemma 17, any k-qubit pure state can be generated by at most G gates. Thus, the space of k-qubit pure states is the same as the space of k-qubit pure states generated by at most G gates. In this way, we can construct a packing net for our subsystem of only the first k qubits by constructing a packing net for all k-qubit pure states. By Theorem 13, there exists an (2ϵ)-packing net $\mathcal{M}_{2\epsilon} = \{\sigma_1, \ldots, \sigma_m\}$ of the space of k-qubit pure-state density matrices satisfying

$$\ln m \geq \frac{2^k}{c}, \quad \chi_0 \leq 4c\epsilon^2 \ln\left(\frac{2^{k-1}}{\epsilon}\right). \tag{B65}$$

From this, we can construct a packing net for our entire n-qubit system as follows:

$$\mathcal{M}'_{2\epsilon} \triangleq \{\sigma_i \otimes |0\rangle\langle 0|^{\otimes(n-k)} : \sigma_i \in \mathcal{M}_{2\epsilon}\}. \tag{B66}$$

We claim that this is indeed a (2ϵ)-packing net of S_1. Let $|\psi\rangle = (U \otimes I)\,|0\rangle^{\otimes n} \in S_1$ and let $\rho = |\psi\rangle\langle\psi|$. Because U only acts on the first k qubits, then we can write $\rho = \rho_k \otimes |0\rangle\langle 0|^{\otimes(n-k)}$, where $\rho_k = U|0\rangle\langle 0|^{\otimes k} U$. Thus, we can see that $\mathcal{M}'_{2\epsilon} \subseteq S_1$. Importantly, all elements of $\mathcal{M}'_{2\epsilon}$ are n-qubit pure states generated by G gates on the first k qubits. Moreover, for any $\sigma'_i, \sigma'_j \in \mathcal{M}'_{2\epsilon}$, we have

$$\begin{aligned} d_{\mathrm{tr}}(\sigma'_i, \sigma'_j) &= d_{\mathrm{tr}}(\sigma_i \otimes |0\rangle\langle 0|^{\otimes(n-k)}, \sigma_j \otimes |0\rangle\langle 0|^{\otimes(n-k)}) \\ &= d_{\mathrm{tr}}(\sigma_i, \sigma_j) > 2\epsilon, \end{aligned} \tag{B67}$$

where the first equality follows by the definition of $\mathcal{M}'_{2\epsilon}$ and the last inequality follows because $\sigma_i, \sigma_j \in \mathcal{M}_{2\epsilon}$.

Hence, $\mathcal{M}'_{2\epsilon}$ is indeed a (2ϵ)-packing net of S_1, which is the set of states we wish to learn. Moreover, it is of the same size as $\mathcal{M}_{2\epsilon}$, which had cardinality m satisfying Eq. (B65). Plugging Eq. (B65) into Theorem 13, we have that in order to learn ρ up to ϵ trace distance, we require

$$\begin{aligned} N_1 &\geq \frac{(1-\delta)\frac{2^k}{c} - \ln 2}{4c\epsilon^2 \ln\left(2^{k-1}/\epsilon\right)} \geq C_1 \frac{(1-\delta)G - C_2}{\epsilon^2 \ln(G/(2\epsilon))} \\ &= \Omega\left(\frac{G(1-\delta)}{\epsilon^2 \log(G/\epsilon)}\right), \end{aligned} \tag{B68}$$

where in the second inequality, C_1 and C_2 are constants, with C_1 depending on c.

This concludes the proof for the second term in the minimum in Proposition 6. Again, for $n < \lfloor \log_2(G/C) \rfloor$, we can appeal to the full quantum state-tomography lower bound of Refs. [21,22]. Thus, we obtain the lower bound

$$N_1 = \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2 \log(G/\epsilon)}\right)\right). \tag{B69}$$

Note, however, that in the limit as $\delta \to 0$ one should find $N \to \infty$. This behavior is not captured in Theorem 13 due to the use of the classical Fano inequality, which treats the measurement procedure as a classical random variable. This behavior is also not present in the lower bounds from Refs. [21,22], where the authors assume that $\delta = \Theta(1)$. In order to recover the dependence on δ, we prove the following lemma.

Lemma 18. Let $|\psi_0\rangle, |\psi_1\rangle$ be any two n-qubit pure quantum states. Suppose that $|\psi_0\rangle$ and $|\psi_1\rangle$ satisfy $d_{\mathrm{tr}}(|\psi_0\rangle, |\psi_1\rangle) \geq \epsilon$. Then, for $\delta \in (0, 1]$,

$$N_2 = \Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right) \tag{B70}$$

copies of $|\psi\rangle \in \{|\psi_0\rangle, |\psi_1\rangle\}$ are needed to distinguish whether $|\psi\rangle = |\psi_0\rangle$ or $|\psi\rangle = |\psi_1\rangle$ with probability at least $1 - \delta$.

Proof. For pure states, we know that the relationship between the fidelity and the trace distance is given by

$$d_{\mathrm{tr}}(|\alpha\rangle, |\beta\rangle) = \sqrt{1 - |\langle\alpha|\beta\rangle|^2}. \tag{B71}$$

In our case, because $d_{\mathrm{tr}}(|\psi_0\rangle, |\psi_1\rangle) \geq \epsilon$, then we have

$$|\langle\psi_0|\psi_1\rangle|^2 \leq 1 - \epsilon^2. \tag{B72}$$

Using the Holevo-Helstrom theorem [142,175], in order to distinguish $|\psi_0\rangle$ from $|\psi_1\rangle$ with probability at least

$1-\delta$, one requires at least N_2 copies of $|\psi\rangle \in \{|\psi_0\rangle, |\psi_1\rangle\}$ satisfying

$$1-\delta \leq \frac{1}{2}+\frac{1}{2}\sqrt{1-|\langle\psi_0|\psi_1\rangle|^{2N_2}}. \tag{B73}$$

Rearranging this inequality, we have

$$N_2 \geq \frac{\log(4\delta(1-\delta))}{\log\left(|\langle\psi_0|\psi_1\rangle|^2\right)} = \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{|\langle\psi_0|\psi_1\rangle|^2}\right)}. \tag{B74}$$

By Eq. (B72), this in particular requires

$$N_2 \geq \frac{\log\left(\frac{1}{4\delta(1-\delta)}\right)}{\log\left(\frac{1}{1-\epsilon^2}\right)} = \Omega\left(\frac{\log(1/\delta)}{\epsilon^2}\right). \tag{B75}$$

■

In our case, note that the conditions of Lemma 18 hold by the existence of the packing net in Eq. (B66), where $|\psi_0\rangle$ and $|\psi_1\rangle$ can be any two states in the packing net. Moreover, because approximating the unknown $|\psi\rangle$ to $(\epsilon/3)$ trace distance suffices to solve the distinguishing task in Lemma 18, then this lower bound also applies for the task of learning a state $|\psi\rangle$. Thus, combining Lemma 18 with Eq. (B69), we have

$$\begin{aligned} N &= \Omega\left(\max\left(N_1, \frac{\log(1/\delta)}{\epsilon^2}\right)\right) \\ &= \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2\log(G/\epsilon)}\right)+\frac{\log(1/\delta)}{\epsilon^2}\right), \end{aligned} \tag{B76}$$

as claimed. ■

This concludes the proof of Proposition 6. Recall that we are seeking a sample-complexity lower bound for states for which we allow our G gates to act on any pair of the n qubits rather than only the first $\lfloor \log_2(G/C) \rfloor$ qubits. We complete the proof of Proposition 5 by reducing to the case of Proposition 6.

Proof of Proposition 5. As before, denote the set of n-qubit quantum states generated by G gates applied to only the first $\lfloor \log_2(G/C) \rfloor$ qubits as S_1. Similarly, denote the set of n-qubit quantum states generated by G gates (applied to any of the qubits) as S_2. Our claim is that the sample complexity of learning states in S_2 is at least the sample complexity of learning states in S_1.

By Proposition 6, we know that the sample complexity of learning states in S_1 is

$$N = \Omega\left(\min\left(\frac{2^n}{\epsilon^2}, \frac{G(1-\delta)}{\epsilon^2\log(G/\epsilon)}\right)+\frac{\log(1/\delta)}{\epsilon^2}\right). \tag{B77}$$

By the definition of sample complexity, this means that there exists some state $\rho \in S_1$ requiring N copies to learn within ϵ trace distance. Then, because $S_1 \subseteq S_2$, then $\rho \in S_2$ as well. Thus, there exists a state $\rho \in S_2$ that requires N copies to learn, so the sample complexity of learning states within S_2 is at least N as well. ■

3. Computational complexity

Theorem 12 states that the sample complexity for learning a description of an unknown n-qubit pure quantum state is linear (up to logarithmic factors) in the number of gates G used to generate the state. Nevertheless, the algorithm described in Appendix B 1 is not computationally efficient, as it constructs and searches over an exponentially large (in G) covering net for all pure states generated by G two-qubit gates. This raises the following question: Does there exist a computationally efficient algorithm?

In this section, we first show that there is no polynomial-time algorithm for learning states generated by $G = \mathcal{O}(n\text{polylog}(n))$ gates, assuming that RingLWE cannot be solved efficiently on a quantum computer. This result also holds for states generated by a depth $d = \mathcal{O}(\text{polylog}(n))$ circuit. Then, we invoke a stronger assumption that RingLWE cannot be solved by any subexponential-time quantum algorithm and show that any quantum algorithm for learning states generated by $\tilde{\mathcal{O}}(G)$ gates must use $\exp(\Omega(G))$ time. This means that the computational hardness already kicks in at $G = \tilde{\omega}(\log n)$. Finally, we explicitly construct an efficient learning algorithm for $G = \mathcal{O}(\log n)$, thus establishing $\log n$ gate complexity as a transition point of computational efficiency. Previous work [56,57] has arrived at similar hardness results for polynomial circuit complexity but our detailed analysis allows us to sharpen the computational lower bound and obtain this transition point.

Theorem 14 (State-learning computational-complexity lower bound assuming polynomial hardness of RingLWE). Let $\lambda = n$ be the security parameter and let $\mathcal{K}$ be the key space parametrized by λ. Let U be a unitary consisting of $G = \mathcal{O}(n\text{polylog}(n))$ gates (or a depth $d = \mathcal{O}(\text{polylog}(n))$ circuit) that prepares a pseudorandom quantum state $|\phi_k\rangle$ for some randomly chosen key $k \in \mathcal{K}$. Such a unitary U exists by Theorem 3, assuming that RingLWE cannot be solved by polynomial-time quantum algorithms. Suppose that we are given $N = \text{poly}(\lambda)$ copies of $|\phi_k\rangle = U|0\rangle^{\otimes n}$. There does not exist a polynomial-time algorithm for learning a circuit description of $|\phi_k\rangle$ to within $\epsilon \leq 1/8$ trace distance with success probability at least $2/3$.

Proof. Suppose, for the sake of contradiction, that there is an efficient algorithm $\mathcal{A}_0$ that can learn a description of $|\phi_k\rangle$ to within ϵ trace distance. Then, by standard boosting of success probability (see, e.g., Ref. [23, Proposition 2.4]), there is an efficient algorithm $\mathcal{A}$ that can learn $|\phi_k\rangle$ to

ALGORITHM 2. Distinguisher $\mathcal{D}$ for PRS.

Input: $\rho^{\otimes N} = |\psi\rangle\langle\psi|^{\otimes N}$
Output: $b \in \{0,1\}$
1 Store one copy of ρ in quantum memory.
2 Run $\mathcal{A}$ on inputs $\rho^{\otimes(N-1)}$, receiving $\hat{\rho}$.
3 Run the SWAP test on the remaining copy ρ and $\hat{\rho}$, receiving a bit $b \in \{0,1\}$.
4 Output b.

the same accuracy with probability at least $p = 1 - 1/128$ with only a constant-factor overhead in time complexity. Note that this boosting requires the distance metric to be efficiently computable, which is guaranteed by the SWAP test elaborated below. We will construct a polynomial-time quantum distinguisher $\mathcal{D}$ that invokes $\mathcal{A}$ to distinguish between $|\phi_k\rangle$ and a Haar-random state $|\phi\rangle$. This contradicts Definition 10.

The distinguisher $\mathcal{D}$ operates according to Algorithm 2.

Recall that the SWAP test [150,151] takes two quantum states σ_1, σ_2 as input and outputs 1 with probability $(1 + \mathrm{tr}(\sigma_1\sigma_2))/2$. We denote this algorithm as $\mathrm{SWAP}(\sigma_1, \sigma_2)$. Note that here we have switched the labels of 0 and 1 compared to the canonical SWAP test presented in Refs. [150,151].

Note that the hypothetical efficient learner $\mathcal{A}$ always produces the circuit description of the output state $\hat{\rho}$ in polynomial time. This means that the circuit description and thus the state $\hat{\rho}$ must also be efficiently implementable. As the SWAP test is also efficient, step 3 of Algorithm 2 can thus indeed be performed efficiently on a quantum computer. Hence, the distinguisher is indeed an efficient quantum algorithm.

Throughout this section, we denote $\rho = |\psi\rangle\langle\psi|$. We analyze the probability that the distinguisher $\mathcal{D}$ outputs 1 when given the pseudorandom state $|\phi_k\rangle$ versus the Haar-random state $|\phi\rangle$.

Case 1: $|\psi\rangle = |\phi_k\rangle$, for a randomly chosen $k \in \mathcal{K}$. We have $\rho = |\psi\rangle\langle\psi| = |\phi_k\rangle\langle\phi_k|$. By the guarantees of $\mathcal{A}$, with probability at least p, we have $d_{\mathrm{tr}}(\hat{\rho}, \rho) \le \epsilon$, where $\hat{\rho}$ is the (potentially mixed) quantum state learned by algorithm $\mathcal{A}$. We can rewrite this as

$$\langle\psi|\hat{\rho}|\psi\rangle \ge 1 - \epsilon, \tag{B78}$$

where we have used the relationship between the fidelity and the trace distance (when one state is pure)

$$d_{\mathrm{tr}}(\rho, \hat{\rho}) \ge 1 - \langle\psi|\hat{\rho}|\psi\rangle. \tag{B79}$$

Then, it immediately follows from Eq. (B78) that

$$\begin{aligned}
&\Pr_{\substack{k\leftarrow\mathcal{K}\\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi_k\rangle^{\otimes N}\right) = 1\right] \\
&= \Pr_{\substack{k\leftarrow\mathcal{K}\\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathrm{SWAP}\left(|\phi_k\rangle\langle\phi_k|, \hat{\rho}\right) = 1\right] \\
&= \mathbb{E}_{k\leftarrow\mathcal{K}}\left[\Pr_{\mathcal{A},\,\mathrm{SWAP}}\left[\mathrm{SWAP}(|\phi_k\rangle\langle\phi_k|, \hat{\rho}) = 1 \,|\, |\phi_k\rangle\right]\right] \\
&\ge p\,\mathbb{E}_{k\leftarrow\mathcal{K}}\left[\frac{1}{2} + \frac{1}{2}(1-\epsilon)\right] = p\left(1 - \frac{\epsilon}{2}\right),
\end{aligned} \tag{B80}$$

where the probability is taken over the random choice of the key $k \in \mathcal{K}$, the randomness in the learning algorithm $\mathcal{A}$ when run on samples $|\phi_k\rangle^{\otimes N}$, and the randomness in the SWAP test. In the inequality, we have split the probability into two terms conditioned on the success and failure of $\mathcal{A}$ and we lower bound the term conditioned on the failure of $\mathcal{A}$ by zero.

Case 2: $|\psi\rangle = |\phi\rangle \sim \mu$, where μ is the Haar measure over pure quantum states. We have $\rho = |\psi\rangle\langle\psi| = |\phi\rangle\langle\phi|$. We want to upper bound the probability that the distinguisher $\mathcal{D}$ outputs 1 when given copies of $|\phi\rangle$. The intuition is that a Haar-random state is likely to be far from any state generated by a circuit with a polynomial-sized description, the space in which the output of $\mathcal{A}$ lies. Let $S_{\mathcal{A}}(|\phi\rangle)$ be the set of quantum states corresponding to all possible outputs of the algorithm $\mathcal{A}$ when run on N copies of $|\phi\rangle$. We follow a similar reasoning as in Eq. (B80) and obtain

$$\begin{aligned}
&\Pr_{\substack{|\phi\rangle\sim\mu\\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi\rangle^{\otimes N}\right) = 1\right] \\
&\le \mathop{\mathbb{E}}_{|\phi\rangle\sim\mu}\left[\max_{\hat{\rho}\in S_{\mathcal{A}}(|\phi\rangle)}\left(\frac{1}{2} + \frac{1}{2}\langle\phi|\hat{\rho}|\phi\rangle\right)\right] + (1-p)
\end{aligned} \tag{B81}$$

$$= \frac{1}{2} + \frac{1}{2}\mathop{\mathbb{E}}_{|\phi\rangle\sim\mu}\left[\max_{\hat{\rho}\in S_{\mathcal{A}}(|\phi\rangle)}\langle\phi|\hat{\rho}|\phi\rangle\right] + (1-p) \tag{B82}$$

$$\triangleq \frac{1}{2} + \frac{1}{2}\mathop{\mathbb{E}}_{|\phi\rangle\sim\mu}[O_\phi] + (1-p), \tag{B83}$$

where in the first line we split the probability according to whether $\mathcal{A}$ succeeds or fails, and we upper bound the failing term by $(1-p)$, and in the last line we define the random variable

$$O_\phi \triangleq \max_{\hat{\rho}\in S_{\mathcal{A}}(|\phi\rangle)}\langle\phi|\hat{\rho}|\phi\rangle. \tag{B84}$$

Furthermore, we can split $\mathbb{E}_{|\phi\rangle\sim\mu}[O_\phi]$ into two parts by introducing a cutoff θ:

$$\begin{aligned}\mathop{\mathbb{E}}_{|\phi\rangle\sim\mu}[O_\phi] &\le \Pr\left[O_\phi \le 1-\frac{\theta}{2}\right]\cdot\left(1-\frac{\theta}{2}\right)\\ &\quad+\Pr\left[O_\phi > 1-\frac{\theta}{2}\right]\cdot 1\\ &\le 1-\frac{\theta}{2}+\Pr\left[O_\phi > 1-\frac{\theta}{2}\right],\end{aligned} \tag{B85}$$

where in the first inequality, we have used that $O_\phi \le 1$. Plugging this into our previous expression, we have

$$\begin{aligned}&\Pr_{\substack{|\phi\rangle\sim\mu\\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi\rangle^{\otimes N}\right)=1\right]\\ &\le 1-\frac{\theta}{4}+\frac{1}{2}\Pr\left[O_\phi > 1.-\frac{\theta}{2}\right]+(1-p)\end{aligned} \tag{B86}$$

We aim to upper bound the probability $\Pr\left[O_\phi > 1-\theta/2\right]$. Note that we have

$$\Pr\left[O_\phi > 1-\frac{\theta}{2}\right] \le \sum_{\hat\rho\in\mathcal{N}_{\sqrt{\theta/2}}}\Pr_{|\phi\rangle\sim\mu}\left[\langle\phi|\,\hat\rho\,|\phi\rangle > 1-\frac{\theta}{2}\right], \tag{B87}$$

where $N_{\sqrt{\theta/2}}$ is a minimal $(\sqrt{\theta/2})$-covering net with respect to the trace distance of the set $S_{\mathcal{A}}(|\phi\rangle)$ of quantum states corresponding to all possible outputs of the algorithm $\mathcal{A}$ when run on N copies of $|\phi\rangle$. We can bound this probability using concentration results. Let $d=2^n$:

$$\Pr_{|\phi\rangle\sim\mu}\left[\langle\phi|\,\hat\rho\,|\phi\rangle > 1-\frac{\theta}{2}\right]$$

$$\le \Pr_{|\phi\rangle\sim\mu}\left[\exp\left(\frac{d}{2}\langle\phi|\,\hat\rho\,|\phi\rangle\right)\ge\exp\left(\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right)\right] \tag{B88}$$

$$\le \exp\left(-\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right)\mathbb{E}_{|\phi\rangle\sim\mu}\left[\exp\left(\frac{d}{2}\langle\phi|\,\hat\rho\,|\phi\rangle\right)\right] \tag{B89}$$

$$= \exp\left(-\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right)\sum_{k=0}^{\infty}\frac{1}{k!}\frac{d^k}{2^k}\mathbb{E}_{|\phi\rangle\sim\mu}[\langle\phi|\,\hat\rho\,|\phi\rangle^k] \tag{B90}$$

$$= \exp\left(-\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right)\sum_{k=0}^{\infty}\frac{1}{k!}\frac{d^k}{2^k}\frac{1}{\binom{k+d-1}{k}}\mathrm{tr}(\hat\rho^{\otimes k}P_{\mathrm{sym}}^{(d,k)}) \tag{B91}$$

$$\le \exp\left(-\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right)\sum_{k=0}^{\infty}\frac{1}{2^k}\mathrm{tr}(\hat\rho^{\otimes k}P_{\mathrm{sym}}^{(d,k)}) \tag{B92}$$

$$\le 2\exp\left(-\frac{d}{2}\left(1-\frac{\theta}{2}\right)\right). \tag{B93}$$

Here, the first two inequalities follow from the following inequality, which holds for $\alpha>0$ and a random variable X:

$$\begin{aligned}\Pr[X\ge\epsilon] &\le \Pr[\exp(\alpha X)\ge\exp(\alpha\epsilon)]\\ &\le \exp(-\alpha X)\mathbb{E}[\exp(\alpha X)].\end{aligned} \tag{B94}$$

The third line follows from the Taylor expansion of $\exp(x)$. The fourth line follows from the identity

$$\mathbb{E}_{|\phi\rangle\sim\mu}\langle\phi|\,O\,|\phi\rangle^k = \frac{1}{\binom{k+d-1}{k}}\mathrm{tr}(O^{\otimes k}P_{\mathrm{sym}}^{(d,k)}), \tag{B95}$$

where we have chosen $O=\hat\rho$ and $P_{\mathrm{sym}}^{(d,k)}$ is the orthogonal projector onto the symmetric subspace of $(\mathbb{C}^d)^{\otimes k}$ (for a proof of this identity, see, e.g., Ref. [157, Example 50]). The fifth line follows from the inequality $1/\binom{k+d-1}{k}\le k!/d^k$. Finally, the last line is true by the following inequalities:

$$\begin{aligned}\mathrm{tr}(\hat\rho^{\otimes k}P_{\mathrm{sym}}^{(d,k)}) &\le \left|\mathrm{tr}(\hat\rho^{\otimes k}P_{\mathrm{sym}}^{(d,k)})\right| &\text{(B96)}\\ &\le \|\hat\rho^{\otimes k}P_{\mathrm{sym}}^{(d,k)}\|_1 &\text{(B97)}\\ &\le \|P_{\mathrm{sym}}^{(d,k)}\|_\infty\|\hat\rho\|_1^k &\text{(B98)}\\ &\le 1, &\text{(B99)}\end{aligned}$$

which follow via properties of the trace norm and because $P_{\mathrm{sym}}^{(d,k)}$ is a projector. Plugging this back into Eq. (B87), we have

$$\begin{aligned}&\Pr\left[O_\phi > 1-\frac{\theta}{2}\right]\\ &\le \sum_{\hat\rho\in\mathcal{N}_{\sqrt{\theta/2}}}\Pr_{|\phi\rangle\sim\mu}\left[\langle\phi|\,\hat\rho\,|\phi\rangle > 1-\frac{\theta}{2}\right] &\text{(B100)}\\ &\le 2\mathcal{N}(S_{\mathcal{A}}(|\phi\rangle), d_{\mathrm{tr}}, \sqrt{\theta/2})\exp\left(-\frac{2^n}{2}\left(1-\frac{\theta}{2}\right)\right). &\text{(B101)}\end{aligned}$$

Moreover, since $S_{\mathcal{A}}(|\phi\rangle)$ is the set of quantum states corresponding to all possible outputs of the algorithm $\mathcal{A}$ when run on $|\phi\rangle^{\otimes N}$, then all states in $S_{\mathcal{A}}(|\phi\rangle)$ must have a poly(n)-size circuit description (because $\mathcal{A}$ is assumed to be efficient). Thus our covering-number upper bound (setting $G=\mathrm{poly}(n)$ in Appendix B 1) implies that

$$\mathcal{N}(S_{\mathcal{A}}(|\phi\rangle), d_{\mathrm{tr}}, \sqrt{\theta/2}) = \mathcal{O}\left((1/\theta)^{\mathrm{poly}(n)}\right). \tag{B102}$$

Thus the above bounds, along with Eq. (B101), give us

$$\Pr\left[O_\phi > 1 - \frac{\theta}{2}\right] = \mathrm{negl}(n), \tag{B103}$$

where $\mathrm{negl}(n)$ denotes a negligible function in n. Putting everything together with Eq. (B86), we have

$$\Pr_{\substack{|\phi\rangle\sim\mu \\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi\rangle^{\otimes N}\right) = 1\right] \le 1 - \frac{\theta}{4} + \mathrm{negl}(n) + (1-p). \tag{B104}$$

Combining with Eq. (B80), we conclude that

$$\left|\Pr_{\substack{k\leftarrow\mathcal{K} \\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi_k\rangle^{\otimes N}\right) = 1\right] - \Pr_{\substack{|\phi\rangle\sim\mu \\ \mathcal{A},\,\mathrm{SWAP}}}\left[\mathcal{D}\left(|\phi\rangle^{\otimes N}\right) = 1\right]\right|$$

$$\ge p\left(1 - \frac{\epsilon}{2}\right) - 2 + \frac{\theta}{4} + p - \mathrm{negl}(n) \tag{B105}$$

$$\ge \frac{1}{16} - \mathrm{negl}(n) \tag{B106}$$

$$\ge \frac{1}{32}, \tag{B107}$$

where we have taken $\theta = 1/2$, $\epsilon \le 1/8$, and $p = 1 - 1/128$, and the last inequality follows by taking n large enough. This contradicts the assumption that $\{|\phi_k\rangle\}_{k\leftarrow\mathcal{K}}$ are pseudorandom quantum states under the assumption that RingLWE cannot be solved by polynomial-time quantum algorithms. ■

Next, we invoke the stronger assumption that RingLWE cannot be solved by any subexponential-time quantum algorithm and show that learning states generated by $\tilde{\mathcal{O}}(G)$ gates require exponential-in-G time.

Theorem 15 (State-learning computational-complexity lower bound assuming subexponential hardness of RingLWE; restatement of lower bound in Theorem 2). Let $\lambda = l = \Theta(G)$, with $l \le n$, be the security parameter and let $\mathcal{K}$ be the key space parametrized by λ. Let U be an l-qubit unitary consisting of $\mathcal{O}(l\mathrm{polylog}(l)) = \mathcal{O}(G\mathrm{polylog}(G))$ gates (or a depth $d = \mathcal{O}(\mathrm{polylog}(G))$ circuit) that prepares an l-qubit pseudorandom quantum state $|\phi_k\rangle$ against subexponential adversaries for some randomly chosen key $k \in \mathcal{K}$. Such a unitary U exists by Theorem 3 assuming that RingLWE cannot be solved by subexponential quantum algorithms. Suppose that we are given $N = \mathrm{poly}(\lambda)$ copies of $|\psi_k\rangle = |\phi_k\rangle \otimes |0\rangle^{\otimes(n-l)} = U|0\rangle^{\otimes n}$. Any quantum algorithm for learning a circuit description of $|\psi_k\rangle$ to within $\epsilon \le 1/8$ trace distance with success probability at least $2/3$ must use $\exp(\Omega(\min\{G, n\}))$ time.

Proof. With the polynomial hardness of RingLWE replaced by subexponential hardness, Theorem 14 asserts that there are no subexponential (in l) quantum algorithms that can learn the l-qubit pseudorandom state $|\phi_k\rangle$ to within trace distance $\epsilon < 1/8$ with success probability at least $2/3$. That is, any such learning algorithms must use at least $\exp(\Omega(l)) = \exp(\Omega(\min\{G, n\}))$ time, since $l \le n$. Meanwhile, a learning algorithm for the n-qubit state $|\psi_k\rangle$ can be used to learn the l-qubit state $|\phi_k\rangle$ in the same run time by postselecting on the last $(n-l)$ qubits being $|0\rangle$, because the trace distance does not increase under such an operation. This implies the $\exp(\Omega(\min\{G, n\}))$-time lower bound for the n-qubit learning algorithm. ■

Finally, we briefly show that learning becomes efficient when $G = \mathcal{O}(\log n)$. The idea is that with $\mathcal{O}(\log n)$ gates, there can only be at most $\mathcal{O}(\log n)$ qubits affected. Thus we can focus on these qubits and learning the states amounts to manipulating vectors of size at most $2^{\mathcal{O}(\log n)} = \mathrm{poly}(n)$, which is efficient. Specifically, we have the following statement.

Proposition 7 (Learning states with logarithmic circuit complexity efficiently; restatement of upper bound in Theorem 2). Let $\epsilon > 0$. Suppose that we are given N copies of a pure n-qubit state $\rho = |\psi\rangle\langle\psi|$, where $|\psi\rangle = U|0\rangle^{\otimes n}$ is generated by a unitary U consisting of $G = \mathcal{O}(\log n)$ two-qubit gates. There exists a learning algorithm that outputs a $\hat{\rho}$ such that $d_{\mathrm{tr}}(\rho, \hat{\rho}) \le \epsilon$ with probability at least $2/3$ using $\mathrm{poly}(n, 1/\epsilon)$ copies and time.

Proof. We prove this by explicitly constructing a learning algorithm based on junta learning (Appendix B 1) and standard tomography methods as follows.

First, we execute Algorithm 1 on copies of ρ and postselect on the trivial qubits being zero as in Appendix B 1. This step uses $\mathrm{poly}(n, 1/\epsilon)$ copies and time and gives us postselected states $\rho' = \rho'' \otimes (|0\rangle\langle 0|)^{\otimes(n-2G)}$ that satisfy $d_{\mathrm{tr}}(\rho, \rho') \le \epsilon/4$ by an appropriate choice of accuracy. Here, ρ'' is a state on $2G = \mathcal{O}(\log n)$ qubits.

Next, we carry out the most straightforward tomography method of measuring all the Pauli coefficients. Concretely, we can represent $\rho'' = \sum_P \alpha_P P$ as a linear combination of all Pauli strings over the $2G$ qubits. Using this representation, we estimate all the coefficients α_P by measuring $\mathrm{tr}(\rho' P)$ and obtain $\hat{\rho} = \hat{\rho}'' \otimes (|0\rangle\langle 0|)^{\otimes(n-2G)}$. By measuring all Pauli-string expectation values $\mathrm{tr}(\rho' P)$ to accuracy $\mathcal{O}(\epsilon/4^{2G})$, we have $d_{\mathrm{tr}}(\rho', \hat{\rho}) \le \epsilon/4$ and thus $d_{\mathrm{tr}}(\rho, \hat{\rho}) \le \epsilon/2$. From standard Chernoff-Hoeffding concentration inequalities, this can be achieved with $\mathcal{O}(4^{2G}/(\epsilon/4^{2G})^2) = \mathrm{poly}(n, 1/\epsilon)$ copies. Finally, we diagonalize $\hat{\rho}''$ and calculate its eigenvector $|\hat{\psi}''\rangle$ with the largest eigenvalue, such that $|\hat{\psi}''\rangle$ is the pure state closest to $\hat{\rho}''$ in trace distance. Let $|\hat{\psi}\rangle = |\hat{\psi}''\rangle \otimes |0\rangle^{\otimes(n-2G)}$. Recall that $d_{\mathrm{tr}}(\rho, \hat{\rho}) \le \epsilon/2$ and ρ is a pure state. Therefore, $d_{\mathrm{tr}}(|\hat{\psi}\rangle\langle\hat{\psi}|, \hat{\rho}) \le d_{\mathrm{tr}}(\rho, \hat{\rho}) \le \epsilon/2$ and thus $d_{\mathrm{tr}}(|\hat{\psi}\rangle\langle\hat{\psi}|, \rho) \le \epsilon$. We output $|\hat{\psi}\rangle$ as the learning outcome the circuit description of which can

be found by finding a unitary with $|\hat{\psi}''\rangle$ as its first column using orthogonalization. Since we are manipulating matrices of size $\mathcal{O}(2^{2G}) = \text{poly}(n)$, the computational complexity is also $\mathcal{O}(n, 1/\epsilon)$. ■

APPENDIX C: LEARNING QUANTUM UNITARIES

In this appendix, we give detailed proofs of Theorem 3 for worst-case unitary learning, of Theorem 4 for average-case unitary learning, and of Theorem 5 for learning with classically described data.

1. Worst-case learning

We begin with the worst-case unitary-learning problem, which measures reconstruction error in terms of the diamond distance $d_\diamond(U, V) = \max_\rho \|(U \otimes I)\rho(U \otimes I)^\dagger - (V \otimes I)\rho(V \otimes I)^\dagger\|_1$. In particular, we consider the task of using queries to an unknown unitary U with bounded circuit complexity G to output a classical circuit description $\hat{U}$ such that $d_\diamond(\hat{U}, U) \leq \epsilon$ with probability at least $2/3$. The diamond distance has a similar operational meaning as the trace distance in state learning. It characterizes the ability to distinguish two processes with arbitrary input states and measurements. If we can learn the unitary with small error in the diamond distance, then we will only make small errors even if we test $\hat{U}$ against U on the worst choice of input states. However, we find the following result stating that this task necessarily requires a number of queries exponential in G, indicating the hardness of worst-case unitary learning.

Theorem 16 (Worst-case unitary learning; restatement of Theorem 3). Given query access to an n-qubit unitary U composed of G two-qubit gates, any algorithm that can output a unitary $\hat{U}$ such that $d_\diamond(\hat{U}, U) \leq \epsilon \in (0, 1/4]$ with probability at least $2/3$ must query U at least $\Omega\left(2^{\min\{G/(2C), n/2\}}/\epsilon\right)$ times, where $C > 0$ is a universal constant. Meanwhile, there exists such an algorithm using $\mathcal{O}(2^n G \log\left(\sqrt{2^n} G/\epsilon\right)/\epsilon)$ queries.

Proof. The upper bound follows from the average-case learning algorithm (Theorem 4, proved below) when working in the exponentially small error regime. Specifically, Theorem 4 gives us an algorithm that uses $\mathcal{O}(G\sqrt{d}\log(G/\epsilon')/\epsilon')$ queries to output a $\hat{U}$ that satisfies $d_{\text{avg}}(\hat{U}, U) \leq \epsilon'$. Meanwhile, from Lemmas 1, 3, and 4, we know that $d_\diamond(\hat{U}, U) \leq 2d_2'(\hat{U}, U) \leq 2\sqrt{d}d_F'(\hat{U}, U) \leq 4\sqrt{d}d_{\text{avg}} \leq 4\sqrt{d}\epsilon'$. Setting $\epsilon = 4\sqrt{d}\epsilon'$, we arrive at the desired worst-case learning query complexity. ■

The proof of the lower bound is inspired by the adversary method [59, Chapter 6] and the optimality of Grover's algorithm [176]. The idea is to construct a set of unitaries that can be distinguished by the worst-case learning algorithm but that only make a minor difference when acting on states, so that a minimal number of queries have to be made in order to distinguish them.

Specifically, we consider all the length-2^k bit strings x that have Hamming weight 1, i.e., $x_i = 1$ for some $i \in [2^k]$ and all the other bits are 0. We focus on the task of distinguishing this set of strings, denoted by X, from the all-zero string $Y = \{0 \ldots 0\}$. We access any such bit strings x through a phase oracle, which is defined as a k-qubit unitary U_x that obeys $U_x|j\rangle = e^{i\epsilon' x_j}|j\rangle$ for all $j \in [2^k]$. In other words, U_x is diagonal and each diagonal element is $e^{i\epsilon'}$ if the corresponding bit is 1 and is 1 if the bit is 0. The unitary for the all-zero string is the identity.

To implement such unitaries with two-qubit gates, we note that since the strings have Hamming weight at most 1, each of the unitaries is equivalent to a $(k-1)$-controlled-phase gate with a proper control rule. The control rule can be realized by $\mathcal{O}(k)$ pairs of one-qubit gates acting on each qubit and the $(k-1)$-controlled-phase gate can be decomposed into $\mathcal{O}(k)$ two-qubit gates [177]. Therefore, with $\mathcal{O}(k)$ gates, one can implement U_x for any 2^k-bit string x with Hamming weight at most 1.

Suppose that Ck gates suffice to implement these U_x. Set $k = \min\{\lfloor G/C \rfloor, n\}$. Then, for any $x \in X \cup Y$, $U_x \otimes I_{n-k}$ is an n-qubit gate composed of at most G gates. Meanwhile, the unitaries for X are far apart from that for Y, because for any $x \in X$—say with $x_j = 1$—we can take another $x' \neq x$ from X with $x'_{j'} = 1$ and let $|\psi_{jj'}\rangle = (|j\rangle + |j'\rangle)/\sqrt{2}$. Then, we have

$$\begin{aligned} & d_\diamond(U_x, U_{0\ldots0}) \\ & \geq \|U_x|\psi_{jj'}\rangle\langle\psi_{jj'}|U_x^\dagger - U_{0\ldots0}|\psi_{jj'}\rangle\langle\psi_{jj'}|U_{0\ldots0}^\dagger\|_1 \\ & = \left\| \frac{e^{i\epsilon'}-1}{2}|j\rangle\langle j'| + \frac{e^{-i\epsilon'}-1}{2}|j'\rangle\langle j| \right\|_1 \\ & = 2\sin\frac{\epsilon'}{2} \geq \frac{\epsilon'}{2}, \end{aligned} \tag{C1}$$

for $\epsilon' \in (0, 1]$. Therefore, if we have a learning algorithm that can learn U_T^n using m queries with accuracy $\epsilon = \epsilon'/4 \in (0, 1/4]$ in the diamond norm with probability $2/3$, it can also distinguish X from Y with the same probability. Note that this also works if the learning algorithm is for (quotient) spectral distance, but not for d_{avg} because $d_{\text{avg}}(U_x, U_{0\ldots0})$ is exponentially small for every x with Hamming weight 1.

In addition, we have the following query-complexity lower bound from the adversary method.

Lemma 19 (Phase-adversary method [59, Lemma 6.4].). Let D be a finite set of functions from a finite set Q to $\mathbb{R}$. To each function $x \in D$, assign an oracle U_x of the form $U_x|q\rangle = e^{ix(q)}|q\rangle$. Let X and Y be two disjoint subsets of D. Let $R \subseteq X \times Y$ be a binary relation on $X \times Y$. For $x \in$

X, we write $R(x) = \{y \in Y : (x,y) \in R\}$ and similarly $R(y)$ for $y \in Y$. Define

$$m = \min_{x\in X} |R(x)|, \quad m' = \min_{y\in Y} |R(y)|,$$
$$l_{q,x} = \sum_{y\in R(x)} |x(q) - y(q)|, \quad l_{q,y} = \sum_{x\in R(y)} |x(q) - y(q)|,$$

and let $l_{\max} = \max_{q\in Q, x\in X, y\in Y} l_{q,x} l_{q,y}$. Then, to distinguish X and Y with success probability at least $2/3$, any algorithm needs at least

$$\Omega\left(\sqrt{\frac{mm'}{l_{\max}}}\right) \tag{C2}$$

queries to the oracle.

For our problem, let $R = X \times Y$. For all bit strings x, define $x(q) = \epsilon x_q$. Then, we have $m = |Y| = 1, m' = |X| = 2^k, l_{q,x} = \epsilon x_q, l_{q,y} = \epsilon$ because for a specific q, only one $x \in R(y) = X$ has $x_q = 1$. Thus $l_{\max} = \epsilon^2$. Plugging these into the above lemma, we obtain a query-complexity lower bound of $\Omega(\sqrt{2^k}/\epsilon)$. Since $k = \min\{\lfloor G/C \rfloor, n\}$, we arrive at the final query-complexity lower bound $\Omega\left(2^{\min\{G/(2C), n/2\}}/\epsilon\right)$.

2. Average-case query-complexity upper bounds

Having seen that worst-case unitary learning is hard, we move on to the setting of average-case learning. In particular, we consider the task of using queries to an unknown unitary U with bounded circuit complexity G to output the classical circuit description of a unitary $\hat{U}$ such that $d_{\text{avg}}(\hat{U}, U) = \sqrt{\mathbb{E}_{|\psi\rangle}[d_{\text{tr}}(\hat{U}|\psi\rangle, U|\psi\rangle)^2]} \le \epsilon$ with probability at least $2/3$. In the following, we give explicit algorithms that solve this learning task with linear-in-G queries, using similar hypothesis-selection techniques as in the state-learning task (Appendix B 1).

Proposition 8 (Average-case unitary-learning upper bounds; upper bounds in Theorem 4). There exists an algorithm that, given query access to an n-qubit unitary U composed of G two-qubit gates, can output a unitary $\hat{U}$ such that $d_{\text{avg}}(\hat{U}, U) \le \epsilon$ with probability at least $2/3$ using

$$\mathcal{O}\left(\min\left\{\frac{4^n}{\epsilon}, \frac{G\log(G/\epsilon)}{\epsilon^2}, \frac{\sqrt{2^n}G\log(G/\epsilon)}{\epsilon}\right\}\right) \tag{C3}$$

queries to the unknown unitary U. Moreover, there is another such algorithm that uses $\mathcal{O}(G\log(G/\epsilon)/\epsilon^4)$ queries without employing auxiliary quantum systems.

The $\mathcal{O}(4^n/\epsilon)$ scaling comes from the diamond-norm learning algorithm in Ref. [23, Theorem 1.1], which directly implies an average-case learning algorithm, because $d_{\text{avg}}(U,V) \le d'_F(U,V) \le d'_2(U,V) \le \frac{1}{\sqrt{2}} d_\diamond(U,V)$, from Lemmas 4, 3, and 1. Note that this part of the bound does not make use of the promise that the unknown unitary can be implemented with G two-qubit gates. In the following, we prove the G-dependent parts of the upper bound.

a. Unitary learning without ancillary systems

We begin by describing the learning algorithm without ancillary systems. The algorithm works similarly to the state-learning procedure. It constructs a covering net over G-gate unitaries with respect to d_{avg} and regards them as candidates for the unknown unitary. In contrast to our state-learning procedure, where the algorithm estimates the trace distance between states, here the algorithm estimates the overlap between unitaries by inputting random states and applying single-shot Clifford classical shadow, which translates into d_{avg}. Then, we select the candidate closest to the unknown unitary as the learning outcome.

Specifically, we consider a $\sqrt{\epsilon'}$-covering net $\mathcal{N}$ of the set of n-qubit unitaries implemented by G two-qubit gates with respect to d_{avg}, as in Corollary 1, and regard the elements $U_i \in \mathcal{N}$ as potential candidates for the unknown unitary U. Our strategy is to use classical shadow to estimate the distances $d_{\text{avg}}(U_i, U)$ for every U_i in the covering net. Then, we can find the one with minimal distance as the output of our learning algorithm.

To achieve this, consider a randomly sampled tensor product of one-qubit stabilizer states,

$$\begin{aligned}|x\rangle &= U_x|0\rangle^{\otimes n} \sim Q \\ &= \text{Uniform}[\{|0\rangle, |1\rangle, |x+\rangle, |x-\rangle, |y+\rangle, |y-\rangle\}^{\otimes n}],\end{aligned} \tag{C4}$$

where $U_x = \otimes_{i=1}^n U_{x_i}$ is the state-preparation unitary and $x \in \mathbb{Z}_6^n$ labels the state. We apply the unknown unitary U to it and obtain $U|x\rangle$. Then, we invoke a single use of the Clifford classical shadow protocol [110]: we randomly sample an n-qubit Clifford gate C and apply it to $U|x\rangle$ and then measure in the computational basis to obtain an outcome $|b\rangle, b \in \{0,1\}^n$, with probability $|\langle b|CU|x\rangle|^2$. Let $\hat{\rho} = (2^n+1)C^\dagger|b\rangle\langle b|C - I$. From Ref. [110], we know that $\mathbb{E}_{C,b}[\hat{\rho}] = U|x\rangle\langle x|U^\dagger$. Now, we consider the observable $O_i = U_i|x\rangle\langle x|U_i^\dagger$ and the estimator $\hat{o}_i = \text{tr}(O_i\hat{\rho})$. Then, we have the expectation value

$$\begin{aligned}\mathop{\mathbb{E}}_{|x\rangle,C,b}[\hat{o}_i] &= \mathop{\mathbb{E}}_{|x\rangle}\left[\text{tr}(O_i \mathop{\mathbb{E}}_{C,b}[\hat{\rho}])\right] \\ &= \mathop{\mathbb{E}}_{|x\rangle}\left[|\langle x|U_i^\dagger U|x\rangle|^2\right] = 1 - d_Q^2(U_i, U),\end{aligned} \tag{C5}$$

where $d_Q(U_i,U) = \sqrt{\mathbb{E}_{|\psi\rangle\sim Q}[d_{\text{tr}}(U_i|\psi\rangle, U|\psi\rangle)^2]}$ is the root-mean-square trace distance with respect to Q as

defined in Lemma 5. Next, we show that $\hat{o}_i$ has bounded variance. Note that

$$\begin{aligned}\mathrm{Var}[\hat{o}_i] &= \mathop{\mathbb{E}}_{|x\rangle,C,b}[\hat{o}_i^2] - \left(\mathop{\mathbb{E}}_{|x\rangle,C,b}[\hat{o}_i]\right)^2 \\ &\le \mathop{\mathbb{E}}_{|x\rangle}\left[\mathop{\mathbb{E}}_{C,b}[\hat{o}_i^2]\right] \le \mathop{\mathbb{E}}_{|x\rangle}[3\mathrm{tr}(O_i^2)] = 3, \end{aligned} \tag{C6}$$

where we have used the variance bound for Clifford shadows [110, Lemma S1 and Proposition S1] and the fact that $\mathrm{tr}(O_i^2) = \mathrm{tr}(O_i) = 1$.

To estimate the expectation values of $\hat{o}_i$, we can draw m independent identically distributed (IID) samples of such input states $\{|x_j\rangle\}_{j=1}^m$ from Q, construct the observables $O_{ij} = U_i |x_j\rangle\langle x_j| U_i^\dagger$, and carry out the above protocol to obtain the estimators $\hat{o}_{ij}$ for $1 \le i \le |\mathcal{N}|, 1 \le j \le m$. Suppose that we take $m = NK$ and construct a median-of-mean estimator

$$\begin{aligned}\hat{o}_i(N,K) &= \mathrm{median}\{\hat{o}_i^{(1)}, \ldots, \hat{o}_i^{(K)}\}, \quad \text{where} \\ \hat{o}_i^{(k)} &= \frac{1}{N}\sum_{j=N(k-1)+1}^{Nk} \hat{o}_{ij}, \quad 1 \le k \le K. \end{aligned} \tag{C7}$$

Then, with the same reasoning as in Ref. [110, Theorem S1], we have the following concentration guarantee: for any $0 < \epsilon', \delta < 1$, if $K = 2\log(2|\mathcal{N}|/\delta)$ and $N = 102/\epsilon'^2$, then

$$|\hat{o}_i(N,K) - (1 - d_Q^2(U_i, U))| \le \epsilon' \quad \text{for all } 1 \le i \le |\mathcal{N}| \tag{C8}$$

with probability at least $1 - \delta$.

With $\hat{o}_i$ in hand, we can select $i^\star \in \mathrm{argmax}_i \hat{o}_i$, and output $U_{i^\star}$. Then, we have

$$\begin{aligned} d_{\mathrm{avg}}(U_{i^\star}, U) &\le \sqrt{2} d_Q(U_{i^\star}, U) \le \sqrt{2(1 - \hat{o}_{i^\star} + \epsilon')} \\ &= \sqrt{2(\epsilon' + \min_i(1 - \hat{o}_i))} \\ &\le \sqrt{2(\epsilon' + \min_i(d_Q^2(U_i, U) + \epsilon'))} \\ &\le \sqrt{8\epsilon'} \end{aligned} \tag{C9}$$

with probability at least $1 - \delta$, where we have used the concentration guarantee in Lemma 5, and $\min_i d_Q^2(U_i, U) \le \min_i 2 d_{\mathrm{avg}}(U_i, U)^2 \le 2\epsilon'$ because $\mathcal{N}$ is a $\sqrt{\epsilon'}$-covering net with respect to d_{avg}. Setting $\epsilon' = \epsilon^2/8$, we arrive at a learning algorithm that uses

$$m = NK = \mathcal{O}(\log(|\mathcal{N}|/\delta)/\epsilon^4) \tag{C10}$$

samples to learn the unknown unitary with accuracy ϵ and success probability at least $1 - \delta$.

If we plug in the covering-number upper bound $\log \mathcal{N} \le \mathcal{O}(G\log(G/\epsilon) + T\log n)$ from Corollary 1, we have sample complexity

$$\mathcal{O}\left(\frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^4}\right) \tag{C11}$$

for large G, say $G \ge n/10$, as desired.

For $G < n/10$, a direct application of the above strategy will give us a suboptimal sample complexity of $\mathcal{O}(G\log(n/\epsilon)/\epsilon^4)$. To overcome this issue, we can carry out a junta learning step similar to Algorithm 1 and Ref. [55] to identify the subset of qubits $A \subset [n]$ on which U acts nontrivially. Since U only has G two-qubit gates, we must have $|A| \le 2G$. The specific procedure is listed in Algorithm 3.

Similarly to Appendix B 1 a, we use Algorithm 3 to identify the nontrivial qubits with high probability. Importantly, from Lemma 11, we have the following guarantee that shows that the expected state on the estimated trivial qubits is close to zero.

Lemma 20. Let $\epsilon, \delta > 0$. Suppose that we are given query access to an n-qubit unitary U composed of G two-qubit gates acting on a subset of the qubits $A \subseteq [n]$. Let $|x\rangle = U_x |0\rangle^{\otimes n}$ be a random tensor product of one-qubit stabilizer states. Let $\rho^x = U_x^\dagger U U_x |0\rangle\langle 0| U_x^\dagger U^\dagger U_x$. Then, Algorithm 3 uses $N = \mathcal{O}\left(G + \log(1/\delta)/\epsilon^2\right)$ queries to U

ALGORITHM 3. Identify qubits acted upon nontrivially (unitary version).

Input: Query access to the unknown unitary U with G two-qubit gates.
Output: List $\hat{A} \subseteq [n]$ of qubits.

1 Initialize $\hat{A} = \emptyset$.

2 Repeat the following $N = \mathcal{O}\left(\frac{G+\log(1/\delta)}{\epsilon^2}\right)$ times:

(a) Sample a random tensor product of 1-qubit stabilizer states $|x\rangle = U_x |0\rangle^{\otimes n}$, apply U and $U_x^\dagger$, and obtain $U_x^\dagger U U_x |0\rangle^{\otimes n}$

(b) Measure in the computational basis and obtain a bit string $|b\rangle, b \in \{0,1\}^n$

(b) Given the measurement outcome $|b\rangle$, set $\hat{A} \leftarrow \hat{A} \cup \mathrm{supp}(b)$, where $\mathrm{supp}(b) = \{i \in [n] : b_i \ne 0\}$.

and outputs, with probability at least $1-\delta$, a list $\hat{A} \subset [n]$ such that

$$\left\langle 0_{\hat{B}} \right| \mathbb{E}_x[\rho_{\hat{B}}^x] \left| 0_{\hat{B}} \right\rangle \geq 1-\epsilon^2, \tag{C12}$$

where $\rho_{\hat{B}}$ denotes the reduced density matrix of ρ when tracing out all qubits other than those in the set $\hat{B} = [n] \setminus \hat{A}$ and $\left|0_{\hat{B}}\right\rangle$ denotes the zero state on all qubits in $\hat{B}$.

Proof. This follows directly from the proof of Lemma 11, because Algorithm 3 is the same as executing Algorithm 1 on the mixed state $\mathbb{E}_x[\rho^x]$, and for the trivial qubits, the $U_x^\dagger$ following U_x and U restores the state to $|0\rangle$. So the proof goes verbatim as in Lemma 11. ■

With this, we can show that ignoring the rest of the qubits $\hat{B} = [n] \setminus \hat{A}$ does not make much of a difference. Let $B = [n] \setminus A$. We again consider a randomly sampled one-qubit stabilizer state and apply U to obtain $|\psi_x\rangle = UU_x |0\rangle^{\otimes n}$. Let $\rho_x = |\psi_x\rangle \langle\psi_x|$ be the associated density matrix and let $U_x^{\hat{B}} = \otimes_{j\in\hat{B}} U_{x_j}$ be the part of U_x that acts on $\hat{B}$. Now, we measure the qubits in $\hat{B}$ in the basis $U_x^{\hat{B}} |b\rangle_{\hat{B}}$, where $b \in \{0,1\}^{|\hat{B}|}$. Note that for qubits in $\hat{B}$, the reduced density matrix in the basis $U_x^{\hat{B}} |b\rangle_{\hat{B}}$ is the same as the $\rho_{\hat{B}}^x$ from Lemma 20 in the junta learning step. So we have $\left\langle 0_{\hat{B}} \right| \mathbb{E}_x[\rho_{\hat{B}}^x] \left| 0_{\hat{B}} \right\rangle \geq 1-\epsilon^2$. After the measurement of the qubits in $\hat{B}$, we do a postselection on the observed measurement outcomes being $U_x^{\hat{B}} |0\rangle_{\hat{B}}$. This postselection is represented by $\Lambda = I_A \otimes (U_x^{\hat{B}} |0\rangle_{\hat{B}} \langle 0| U_x^{\hat{B}\dagger})$, with $\Lambda^2 = \Lambda$. Let $\rho_x' = \sqrt{\Lambda}\rho_x\sqrt{\Lambda}/\mathrm{tr}(\Lambda\rho_x)$ be the postselected state. Now, we want to show ρ_x' is close to ρ_x on average. We invoke the following gentle measurement lemma for normalized ensembles.

Lemma 21 (Gentle measurement lemma for normalized ensembles; variant of Ref. [155, Lemma 9.4.3]). Let $\{x, \rho_x\}$ be an ensemble of states. If Λ is a positive-semidefinite operator with $\Lambda \leq I$ and $\mathrm{tr}(\Lambda \mathbb{E}_x[\rho_x]) \geq 1-\epsilon$, where $\epsilon \in [0,1]$, then

$$\mathbb{E}_x \left\| \rho_x - \frac{\sqrt{\Lambda}\rho_x\sqrt{\Lambda}}{\mathrm{tr}(\Lambda\rho_x)} \right\|_1 \leq 3\sqrt{\epsilon}. \tag{C13}$$

Proof. Let $\rho_x' = \sqrt{\Lambda}\rho_x\sqrt{\Lambda}/\mathrm{tr}(\Lambda\rho_x)$. From Ref. [155, Lemma 9.4.3], we know that $\mathbb{E}_x \left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1 \leq 2\sqrt{\epsilon}$. Note that the left-hand side can be lower bounded by

$$\begin{aligned}
&\mathbb{E}_x \left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1 \\
&= \mathbb{E}_x \left\| \rho_x - \rho_x' + \rho_x' - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1 \\
&\geq \mathbb{E}_x \|\rho_x - \rho_x'\|_1 - \mathbb{E}_x \|\rho_x' - \sqrt{\Lambda}\rho_x\sqrt{\Lambda}\|_1 \\
&= \mathbb{E}_x \|\rho_x - \rho_x'\|_1 - \mathbb{E}_x (1 - \mathrm{tr}(\Lambda\rho_x)) \|\rho_x'\|_1 \\
&\geq \mathbb{E}_x \|\rho_x - \rho_x'\|_1 - \epsilon,
\end{aligned} \tag{C14}$$

where we have used the triangle inequality, $\|\rho_x'\|_1 = 1$, and $\mathrm{tr}(\Lambda \mathbb{E}_x[\rho_x]) \geq 1-\epsilon$. Therefore, we arrive at

$$\mathbb{E}_x \|\rho_x - \rho_x'\|_1 \leq 2\sqrt{\epsilon} + \epsilon \leq 3\sqrt{\epsilon}, \tag{C15}$$

because $\epsilon \in [0,1]$, concluding the proof of Lemma 21. ■

Using Lemma 21 for our scenario, we have $\mathbb{E}_x \|\rho_x - \rho_x'\|_1 \leq 3\epsilon$ with probability at least $1-\delta$. After the postselection, we apply the same Clifford-shadow strategy as in the $T \geq n/10$ case, with two differences. First, note that after postselection, the action on every qubit in $\hat{B}$ is identity. So we can, without loss of generality, pick an arbitrary subset A' of those qubits in $\hat{B}$ as $A \setminus \hat{A}$ and consider an $\sqrt{\epsilon}$-covering net $\mathcal{N}$ of G gate unitaries on qubits $\hat{A} \cup A'$ with respect to d_{avg}, with $|\hat{A} \cup A'| = |A| \leq 2G$. Then, we have $\min_{U_i \in \mathcal{N}} d_{\mathrm{avg}}(U_i, U) \leq \epsilon$, and $\log|\mathcal{N}| \leq \mathcal{O}(G\log(G/\epsilon) + G\log(|A \cup A'|)) \leq \mathcal{O}(G\log(G/\epsilon))$. Second, for each element U_i in the covering net, we can construct an observable $O_i = U_i |x\rangle \langle x| U_i^\dagger$ similar to before but now the estimator will concentrate around a slightly different expectation value. Specifically, if we use a median-of-means estimator $\hat{o}_i(N,K)$ with $K = 2\log(2|\mathcal{N}|/\delta)$ and $N = 102/\epsilon^2$, then we have

$$|\hat{o}_i(N,K) - \mathbb{E}_x[\mathrm{tr}(\rho_x' O_i)]| \leq \epsilon \quad \text{for all } 1 \leq i \leq |\mathcal{N}| \tag{C16}$$

with probability at least $1-\delta$. Nevertheless, since ρ_x' and ρ_x are close on average, we have

$$\begin{aligned}
&|\hat{o}_i(N,K) - (1 - d_Q^2(U_i \otimes I, U))| \\
&= |\hat{o}_i(N,K) - \mathbb{E}_x[\mathrm{tr}(\rho_x' O_i)] \\
&\quad + \mathbb{E}_x[\mathrm{tr}(\rho_x' O_i)] - \mathbb{E}_x[\mathrm{tr}(\rho_x O_i)]| \\
&\leq |\hat{o}_i(N,K) - \mathbb{E}_x[\mathrm{tr}(\rho_x' O_i)]| \\
&\quad + \mathbb{E}_x[|\mathrm{tr}(\rho_x' O_i) - \mathrm{tr}(\rho_x O_i)|] \\
&\leq \epsilon + \mathbb{E}_x[\|\rho_x' - \rho_x\|_1 \|O_i\|] \\
&\leq \epsilon + 3\epsilon = 4\epsilon, \quad \text{for all } 1 \leq i \leq |\mathcal{N}|
\end{aligned} \tag{C17}$$

with probability at least $1-2\delta$, where we have used the triangle inequality, $\|O_i\| = 1$, and $\mathbb{E}_x \|\rho_x - \rho_x'\|_1 \leq 3\epsilon$. With

this concentration guarantee, we can select the candidate with the largest $\hat{o}_i$: $i^\star \in \mathrm{argmax}_i \hat{o}_i$ and output $U_{i^\star} \otimes I$. As before, we have, with probability at least $1 - 2\delta$,

$$\begin{aligned} d_{\mathrm{avg}}(U_{i^\star} \otimes I, U) &\le \sqrt{2} d_Q(U_{i^\star} \otimes I, U) \\ &\le \sqrt{2(4\epsilon + 2\epsilon + 4\epsilon)} = \sqrt{20\epsilon}. \end{aligned} \tag{C18}$$

Redefining 20ϵ to be ϵ^2 and 2δ to be δ, we arrive at a learning algorithm that uses

$$\begin{aligned} m &= \mathcal{O}\left(\frac{G + \log(1/\delta)}{\epsilon^4}\right) + NK \\ &= \mathcal{O}\left(\frac{G\log(G/\epsilon) + \log(1/\delta)}{\epsilon^4}\right) \end{aligned} \tag{C19}$$

queries to the unitary to learn it with accuracy ϵ in d_{avg} and success probability at least $1 - \delta$ when $G < n/10$. Combined with the case of $G \ge n/10$, this concludes the learning algorithm without an ancillary system in Proposition 8.

b. Unitary learning with ancillary systems

The above $\mathcal{O}(1/\epsilon^4)$ scaling is suboptimal. It arises from the fact that in the classical shadow estimation, the estimated quantity is the square of d_{avg} rather than d_{avg} itself. To improve the ϵ dependence, we make use of ancillary systems via the Choi-Jamiołkowski duality [63–65]. Specifically, we consider the maximally entangled state over a pair of n-qubit systems $|\Phi\rangle = 1/\sqrt{d}\sum_{i=1}^{2^n} |i\rangle \otimes |i\rangle$ and define the Choi state $|U\rangle\!\rangle$ corresponding to a unitary U as $|U\rangle\!\rangle = (U \otimes I)|\Phi\rangle$. That is, the Choi state $|U\rangle\!\rangle$ of an n-qubit unitary U is a pure $(2n)$-qubit state constructed by applying U on half of the qubits in n EPR pairs. For any subset $A \subseteq [n]$ of the qubits that are acted upon by U, we refer to the corresponding $|A|$ qubits in the EPR pairs as the entangled qubits corresponding to A. We note the following fact, which relates the trace distance between Choi states to the average-case distance between the unitaries.

Lemma 22 (Equivalence of trace distance between Choi states and average-case distance). Let $U, V \in U(2^n)$ be two n-qubit unitaries, let $|\Phi\rangle = 1/\sqrt{d}\sum_{i=1}^{2^n} |i\rangle \otimes |i\rangle$ be a maximally entangled state, and let $|U\rangle\!\rangle = (U \otimes I)|\Phi\rangle, |V\rangle\!\rangle = (V \otimes I)|\Phi\rangle$ be the corresponding Choi states. Then, we have

$$\frac{1}{\sqrt{2}} d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle) \le d_{\mathrm{avg}}(U, V) \le d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle). \tag{C20}$$

Proof. By the standard conversion between the fidelity and the trace distance between pure states, we have

$$d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle) = \sqrt{1 - |\langle\!\langle U|V\rangle\!\rangle|^2} = \sqrt{1 - \frac{1}{d^2}|\mathrm{tr}(U^\dagger V)|^2}, \tag{C21}$$

where the last step uses that $\langle\Phi| A \otimes B |\Phi\rangle = \frac{1}{d}\mathrm{tr}[A^T B]$ (cf., e.g., Ref. [178, Example 1.2]). On the other hand, from Eq. (A14), we have

$$d_{\mathrm{avg}}(U, V) = \sqrt{1 - \frac{d + |\mathrm{tr}(U^\dagger V)|^2}{d^2 + d}}. \tag{C22}$$

Combining these two equations, we obtain

$$\begin{aligned} d_{\mathrm{avg}}(U, V) &= \sqrt{\frac{d}{d+1}} d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle) \\ &\in \left[\frac{1}{\sqrt{2}} d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle), d_{\mathrm{tr}}(|U\rangle\!\rangle, |V\rangle\!\rangle)\right]. \end{aligned} \tag{C23}$$

■

With Lemma 22, we construct a covering net over Choi states corresponding to G-gate unitaries as follows. From Corollary 1, we take an ϵ'-covering net $\mathcal{N}$ of G-gate unitaries with respect to d_{avg} that has cardinality $|\mathcal{N}| \le \mathcal{O}(G\log(G/\epsilon) + G\log n)$. Then, for any G-gate unitary U, there exists a $U_i \in \mathcal{N}$ such that $d_{\mathrm{avg}}(U, U_i) \le \epsilon'$. Hence $d_{\mathrm{tr}}(|U\rangle\!\rangle, |U_i\rangle\!\rangle) \le \sqrt{2} d_{\mathrm{avg}}(U, U_i) \le \sqrt{2}\epsilon'$ by Lemma 22. Therefore, the Choi states of the unitaries in $\mathcal{N}$ form a $(\sqrt{2}\epsilon)$-covering net of the Choi states of G-gate unitaries.

Now, we can use these pure Choi states as candidates for hypothesis selection. By Proposition 1, the hypothesis-selection algorithm based on classical shadow uses $\mathcal{O}(\log(|\mathcal{N}|/\delta)/\epsilon'^2)$ samples of the Choi state $|U\rangle\!\rangle$ to output a candidate $|\hat{U}\rangle\!\rangle$, $\hat{U} \in \mathcal{N}$, such that $d_{\mathrm{tr}}(|U\rangle\!\rangle, |\hat{U}\rangle\!\rangle) \le 3\sqrt{2}\epsilon' + \epsilon'$ with probability at least $1 - \delta$. Setting $(3\sqrt{2} + 1)\epsilon' = \epsilon$, we find a $\hat{U}$ such that $d_{\mathrm{avg}}(\hat{U}, U) \le d_{\mathrm{tr}}(|U\rangle\!\rangle, |\hat{U}\rangle\!\rangle) \le \epsilon$ with probability at least $1 - \delta$ using

$$\mathcal{O}\left(\frac{G\log(G/\epsilon) + G\log n + \log(1/\delta)}{\epsilon^2}\right) \tag{C24}$$

queries to the unknown unitary U. When $G \ge n/10$, this gives the desired $\mathcal{O}((G\log(G/\epsilon) + \log(1/\delta))/\epsilon^2)$ query complexity.

For $G < n/10$, we again need a junta learning step to identify the set of qubits $A \subseteq [n]$ that are acted on nontrivially. To do this, we follow the idea of Algorithms 1 and 3 and Ref. [55, Algorithm 8] and consider Algorithm 4 that makes use of Choi states of Pauli matrices $\sigma_0 = I, \sigma_1 = X, \sigma_2 = Y, \sigma_3 = Z$.

ALGORITHM 4. Identify qubits acted upon nontrivially (Choi version).

Input: Query access to the unknown unitary U with G two-qubit gates.
Output: List $\hat{A} \subseteq [n]$ of qubits.
1 Initialize $\hat{A} = \emptyset$.
2 Repeat the following $N = \mathcal{O}\left(\frac{G+\log(1/\delta)}{\epsilon^2}\right)$ times:
(a) Prepare the Choi state $|U\rangle\!\rangle$ by applying $U \otimes I$ to the maximally entangled state $|\Phi\rangle$
(b) Measure in the basis of Pauli Choi states $|\sigma_x\rangle\!\rangle = \otimes_{i=1}^{n}(\sigma_{x_i} \otimes I)\,|\Phi\rangle\,, x \in \mathbb{Z}_4^n$, and obtain a string $|b\rangle\,, b \in \{0,1,2,3\}^n$
(b) Given the measurement outcome $|b\rangle$, set $\hat{A} \leftarrow \hat{A} \cup \mathrm{supp}(b)$, where $\mathrm{supp}(b) = \{i \in [n] : b_i \neq 0\}$.

Similarly to Lemmas 11 and 20, we have the following guarantee that the Choi state on the estimated trivial qubits is close to the Choi state of the identity.

Lemma 23. Let $\epsilon, \delta > 0$. Suppose that we are given query access to an n-qubit unitary U composed of G two-qubit gates acting on a subset of the qubits $A \subseteq [n]$. Let $\rho = |U\rangle\!\rangle\langle\!\langle U|$ be the Choi state of U. Then, Algorithm 4 uses $N = \mathcal{O}\left((G + \log(1/\delta))/\epsilon^2\right)$ queries to U and outputs, with probability at least $1 - \delta$, a list $\hat{A} \subset [n]$ such that

$$\langle\!\langle I_{\hat{B}}|\rho_{\hat{B}}|I_{\hat{B}}\rangle\!\rangle \geq 1 - \epsilon^2, \tag{C25}$$

where $\rho_{\hat{B}}$ denotes the reduced density matrix for ρ by tracing out all qubits other than those in the set $\hat{B} = [n] \setminus \hat{A}$ and the corresponding entangled qubits and $|I_{\hat{B}}\rangle\!\rangle$ denotes the Choi state of the identity on qubits in $\hat{B}$.

Proof. The proof goes similarly to that of Lemma 11, except that $|0\rangle$ is replaced by $|I\rangle\!\rangle$. The measurement over Pauli Choi states in Algorithm 4 can be understood as measuring each entangled pair of qubits in the basis $\{|I\rangle\!\rangle, |X\rangle\!\rangle, |Y\rangle\!\rangle, |Z\rangle\!\rangle\}$ and gives an element from $\{0,1,2,3\} = \mathbb{Z}_4$. Specifically, let A' be any set that could be output by Algorithm 4. We want to identify A' with the actual identified set $\hat{A}$. Let $B' \triangleq [n] \setminus A'$. Let $E_{i,A'}$ be the event that round i of measurement of the qubits in $B' = [n] \setminus A'$ in Algorithm 4 yields the all-zero $\mathbb{Z}_4$ string. Let $X_{i,A'}$ be the indicator random variable corresponding to the event $E_{i,A'}$. Then, we have that $\bar{X}_{A'} \triangleq 1/N \sum_{i=1}^{N} X_{i,A'}$ is the number of times that the entangled pairs in B' are all measured to be zero divided by the total number of measurements. In other words, $\bar{X}_{A'}$ is the estimated overlap that the state $\rho_{B'}$ on qubits in B' has with the identity Choi state on B'. Moreover, we have

$$\mathbb{E}[X_{A'}] \triangleq \mathbb{E}[X_{i,A'}] = \langle\!\langle I_{B'}|\rho_{B'}|I_{B'}\rangle\!\rangle, \tag{C26}$$

for all A'. This says that the true expectation of our random variables is the true overlap of the state $\rho_{B'}$ with the identity Choi state on B'. Then, we have the same Claim 1 as in Lemma 11 and so Lemma 23 follows. ■

With this, we can again show that ignoring the rest of the qubits $\hat{B} = [n] \setminus \hat{A}$ does not make much difference. Let $B = [n] \setminus A$. We prepare the Choi state $\rho = |U\rangle\!\rangle\langle\!\langle U|, |U\rangle\!\rangle = (U \otimes I)\,|\Phi\rangle$ and measure in the basis of Pauli Choi states over the qubits in $\hat{B}$: $\{|\sigma_x\rangle\!\rangle_{\hat{B}} : x \in \mathbb{Z}_4^{|\hat{B}|}\}$. After the measurement, we do a postselection on the observed measurement outcomes being $|I_{\hat{B}}\rangle\!\rangle$. This postselection is represented by $\Lambda = I \otimes |I_{\hat{B}}\rangle\!\rangle\langle\!\langle I_{\hat{B}}|$, with $\Lambda^2 = \Lambda$, and the first identity over the entangled pairs outside $\hat{B}$. Let $\rho' = \sqrt{\Lambda}\rho\sqrt{\Lambda}/\mathrm{tr}(\Lambda\rho)$ be the postselected state. Now, we want to show that ρ' is close to ρ. From Lemma 23, we know that $\mathrm{tr}(\Lambda\rho) \geq 1 - \epsilon^2$ with probability at least $1 - \delta$. Then, by the gentle measurement lemma (Lemma 13), we have $d_{\mathrm{tr}}(\rho', \rho) \leq \epsilon$ with the same probability.

Now, we can apply the hypothesis-selection protocol to ρ' as in the $G > n/10$ case but with a different covering net. Specifically, note that after postselection, the action on every entangled pair in $\hat{B}$ is identity. So we can, with loss of generality, pick an arbitrary subset A' of those qubits in $\hat{B}$ as $A \setminus \hat{A}$ and consider an ϵ-covering net $\mathcal{N}$ of G gate unitaries on qubits $\hat{A} \cup A'$ with respect to d_{avg}, with $|\hat{A} \cup A'| = |A| \leq 2G$, with each element tensor product with identity over the rest of the qubits. Then, we have $\min_{U_i \in \mathcal{N}} d_{\mathrm{tr}}(|U_i\rangle\!\rangle, |U\rangle\!\rangle) \leq \sqrt{2} d_{\mathrm{avg}}(U_i, U) \leq \sqrt{2}\epsilon$, and $\log|\mathcal{N}| \leq \mathcal{O}(G\log(G/\epsilon) + G\log(|A \cup A'|)) \leq \mathcal{O}(G\log(G/\epsilon))$. Since $d_{\mathrm{tr}}(\rho', \rho) \leq \epsilon$, we also have $\min_{U_i \in \mathcal{N}} d_{\mathrm{tr}}(|U_i\rangle\!\rangle\langle\!\langle U_i|, \rho') \leq \epsilon + \sqrt{2}\epsilon = (\sqrt{2}+1)\epsilon$.

With this covering net, we apply the hypothesis selection based on classical shadow (Proposition 1) to ρ'. This procedure uses $\mathcal{O}(\log(|\mathcal{N}|/\delta)/\epsilon^2)$ copies of ρ' (each prepared using one query to U) and outputs a $|U_{i^\star}\rangle\!\rangle$ such that $d_{\mathrm{tr}}(|U_{i^\star}\rangle\!\rangle\langle\!\langle U_{i^\star}|, \rho') \leq 3(\sqrt{2}+1)\epsilon + \epsilon = (3\sqrt{2}+4)\epsilon$ with probability at least $1 - \delta$. This means that

$$\begin{aligned} d_{\mathrm{avg}}(U_{i^\star}, U) &\leq d_{\mathrm{tr}}(|U_{i^\star}\rangle\!\rangle, |U\rangle\!\rangle) \\ &\leq d_{\mathrm{tr}}(|U_{i^\star}\rangle\!\rangle\langle\!\langle U_{i^\star}|, \rho') + d_{\mathrm{tr}}(\rho', \rho) \\ &\leq (3\sqrt{2}+4)\epsilon + \epsilon = 4(\sqrt{2}+1)\epsilon, \end{aligned} \tag{C27}$$

with a total probability at least $1 - 2\delta$ (considering both the junta learning and hypothesis selection).

Therefore, by redefining $4(\sqrt{2}+1)\epsilon$ to be ϵ and 2δ to be δ, we arrive at a desired algorithm for $G \leq n/10$ that

uses, in total,

$$\mathcal{O}\left(\frac{G+\log(1/\delta)}{\epsilon^2}\right)+\mathcal{O}\left(\frac{G\log(G/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)=\mathcal{O}\left(\frac{G\log(G/\epsilon)+\log(1/\delta)}{\epsilon^2}\right) \tag{C28}$$

queries to the unknown unitary U. Combined with the $G \geq n/10$ case, we conclude the learning algorithm with ancillary systems that achieves the $\mathcal{O}((G\log(G/\epsilon)+\log(1/\delta)/\epsilon^2)$ query complexity in Proposition 8.

c. Bootstrap to improve ϵ dependence

To further improve the ϵ dependence, we modify the bootstrap method in Ref. [23] and achieve a Heisenberg scaling $\tilde{\mathcal{O}}(1/\epsilon)$. However, with our average-case distance, which can only control the average behavior of the eigenvalues of the unitaries, we are unable to perform the bootstrap for general ϵ. Instead, the bootstrap works only when the error is exponentially small, $\epsilon = O(1/\sqrt{d})$, and achieves the Heisenberg scaling at the cost of a dimensional factor, leading to a query complexity of

$$\mathcal{O}\left(\frac{\sqrt{2^n}(G\log(G/\epsilon)+\log(1/\delta))}{\epsilon}\right). \tag{C29}$$

Whether a general Heisenberg scaling without dimension-dependent scaling is achievable remains open.

Now, we state the bootstrap method in Algorithm 5, which uses the unitary-learning algorithm with ancillary systems (Appendix C 2 b) as a subroutine. We need to prove two things about Algorithm 5: (1) that it outputs a $\hat{U}$ that satisfies $d_{\text{avg}}(\hat{U}, U) \leq \epsilon$ with probability at least $1-\delta$; and (2) that the query complexity is $\mathcal{O}\left(\sqrt{d}(G\log(G/\epsilon)+\log(1/\delta))/\epsilon\right)$.

We first prove (1) by induction. Before doing so, we need to show that the learning algorithm $\mathcal{A}$ can indeed learn $(UV_j^\dagger)^{p_j}$ well for all j. Let $c = 10^{-5}$. Note that with the definition of $\mathcal{N}$, we know that for any G-gate unitary U, $\exists U_i \in \mathcal{N}$ such that $d_{\text{avg}}(U, U_i) \leq c\epsilon$, and therefore

$$\begin{aligned} d_{\text{avg}}((U_iV_j^\dagger)^{p_j}, (UV_j^\dagger)^{p_j}) &\leq d'_F((U_iV_j^\dagger)^{p_j}, (UV_j^\dagger)^{p_j}) \\ &\leq p_j d'_F(U_i, U) \leq 2p_j d_{\text{avg}}(U_i, U) \\ &\leq 4c/\sqrt{d}, \end{aligned} \tag{C30}$$

where we have used items (1) and (2) in Lemma 4, unitary invariance of d'_F, and $p_j = 2^j \leq 2^t \leq 2/(\epsilon\sqrt{d})$. Thus $\{(U_iV_j^\dagger)^{p_j}, U_i \in \mathcal{N}\}$ forms a $4c/\sqrt{d}$-covering net of $\{(UV_j^\dagger)^{p_j} \mid U \text{ is a } G\text{-gate unitary}\}$, which can be used by the hypothesis-selection algorithm $\mathcal{A}$ as a set of candidates. The output R_j of $\mathcal{A}$ satisfies $d'_F(R_j, (UV_j^\dagger)^{p_j}) \leq 2d_{\text{avg}}(R_j, (UV_j^\dagger)^{p_j}) \leq 4(\sqrt{2}+1)\cdot 4c/\sqrt{d} < 40c/\sqrt{d}$ (see Eq. (C27)). The number of queries to U that this procedure uses is $\mathcal{O}\left(p_j(G\log(G/c\epsilon)+\log(1/\eta_j))/(4c/\sqrt{d})^2\right) = \mathcal{O}\left(p_j d(G\log(G/\epsilon)+\log(1/\eta_j))\right)$.

Now, we proceed to prove (1) by induction. Let us assume that the learning algorithm succeeds for all $j = 1, \ldots, t$. Let $\delta_j = d'_F(U, V_j) = d'_F(UV_j^\dagger, I)$ be the error after iteration $j-1$. We will prove that $\delta_k \leq 2^{-k-5}/\sqrt{d}$. For iteration 0, we have $p_0 = 1$ and by the accuracy of $\mathcal{A}$, we know that $\delta_1 = d'_F(U, V_1) < 40c/\sqrt{d} < 2^{-6}/\sqrt{d}$. Now, we assume that $\delta_k \leq 2^{-k-5}/\sqrt{d}$ and prove that $\delta_{k+1} \leq 2^{-k-6}/\sqrt{d}$. Note that $(UV_k^\dagger)^{p_k}$ and R_k are sufficiently close to identity in the sense that

$$\begin{aligned} d'_F((UV_k^\dagger)^{p_k}, I) &\leq p_k d'_F(UV_k^\dagger, I) = p_k\delta_k \\ &\leq \frac{2^{-5}}{\sqrt{d}} < \frac{4/(25\pi)}{\sqrt{d}} \end{aligned} \tag{C31}$$

ALGORITHM 5. Bootstrapping to Heisenberg scaling.

Input: Query access to the unknown n-qubit G-gate unitary U.
An error parameter $\epsilon \in (0, 1/\sqrt{d})$.
Output: A unitary $\hat{U}$.
Let $t \leftarrow \left\lceil \log_2(1/(\epsilon\sqrt{d})) \right\rceil$.
Let $V_0 \leftarrow I$.
Let $\mathcal{N} \leftarrow$ an $(\epsilon/10^5)$-covering net of G-gate unitaries with respect to d_{avg}.
for $j \leftarrow 0$ *to* t **do**
  Let $p_j \leftarrow 2^j$.
  Let $\eta_j \leftarrow 8^{j-t-1}\delta$.
  Use the algorithm $\mathcal{A}$ in Appendix C 2 b with success probability $1-\eta_j$ and accuracy $1/(25000\sqrt{d})$ to find a candidate R_j in $\{(U_iV_j^\dagger)^{p_j} \mid U_i \in \mathcal{N}\}$ that is closest to $(UV_j^\dagger)^{p_j}$ in d_{avg}.
  Let $V_{j+1} \leftarrow R_j^{1/p_j}V_j$.
return $\hat{U} \leftarrow V_{t+1}$

and

$$d'_F(R_k, I) \le d'_F(R_k, (UV_k^\dagger)^{p_k}) + d'_F((UV_k^\dagger)^{p_k}, I) \le \frac{40c}{\sqrt{d}} + \frac{2^{-5}}{\sqrt{d}} < \frac{4/(25\pi)}{\sqrt{d}}. \tag{C32}$$

Thus, we can invoke item (3) of Lemma 4 and obtain

$$\delta_{k+1} = d'_F(U, V_{k+1}) = d'_F(UV_k^\dagger, R_k^{1/p_k}) \le \frac{2}{p_k} d'_F((UV_k^\dagger)^{p_k}, R_k) \le \frac{80c}{p_k\sqrt{d}} < \frac{2^{-k-6}}{\sqrt{d}}. \tag{C33}$$

Therefore, by induction, we have shown that $\delta_k \le 2^{-k-5}/\sqrt{d}$. At the end of the iteration, when $k = t = \left\lceil \log_2(1/(\sqrt{d}\epsilon)) \right\rceil$, we have

$$\delta_{t+1} = d'_F(U, V_{t+1}) \le \frac{2^{-t-6}}{\sqrt{d}} < \epsilon. \tag{C34}$$

The above accuracy is conditioned on the success of all executions of the learning algorithm. By the union bound, the failure probability is upper bounded by

$$\sum_{j=0}^{t} \eta_j = \delta \sum_{j=0}^{t} 8^{-(t-j)-1} = \delta \sum_{j=0}^{t} 8^{-j-1} < \delta. \tag{C35}$$

This concludes the proof of (1).

Next, we move on to (2) and count the overall number of queries to the unknown unitary. Summing over all iterations, the number of queries is

$$\mathcal{O}\left(\sum_{j=0}^{t} p_j\, d(G\log(G/\epsilon) + \log(1/\eta_j))\right) = \mathcal{O}\left(dG\log(G/\epsilon)\sum_{j=0}^{t} 2^j + d\log(1/\delta)\sum_{j=0}^{t} 2^j\,(t-j+1)\right) = \mathcal{O}\left(d(G\log(G/\epsilon) + \log(1/\delta))2^t\right) = \mathcal{O}\left(\frac{\sqrt{d}(G\log(G/\epsilon) + \log(1/\delta))}{\epsilon}\right). \tag{C36}$$

This concludes the proof of the $\mathcal{O}(1/\epsilon)$ scaling algorithm in Proposition 8.

Finally, we note that an analogous bootstrap method can also be applied to improve the ϵ dependence for our unitary-learning procedure without auxiliary systems, albeit again incurring a dimension factor. Namely, a variant of Algorithm 5 relying on the algorithm of Appendix C 2 a as a subroutine succeeds at outputting a $\hat{U}$ that satisfies $d_{\mathrm{avg}}(\hat{U}, U) \le \epsilon$ with probability at least $1-\delta$ using $\mathcal{O}\left(d^{3/2}(G\log(G/\epsilon) + \log(1/\delta))/\epsilon\right)$ queries to the unknown unitary U, assuming that $\epsilon < 1/d^{3/2}$.

3. Average-case query-complexity lower bounds

For the lower bound, we construct a packing net consisting of G-gate unitaries that are pairwise sufficiently far apart that an average-case learning algorithm can discriminate them. Meanwhile, the success probability of distinguishing a set of unitaries is upper bounded by the number of queries made [68]. This gives us an $\Omega(G)$ query-complexity lower bound. To incorporate ϵ dependence, we follow [23] and map the problem into a fractional-query problem [66,67]. In this way, we arrive at the following result.

Proposition 9 (Average-case unitary-learning lower bound; lower bound in Theorem 4). Let U be an n-qubit unitary composed of G two-qubit gates. Any algorithm that, given query access to U, $U^\dagger$, $cU = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes U$ and $cU^\dagger = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes U^\dagger$, can output a unitary $\hat{U}$ such that $d_{\mathrm{avg}}(\hat{U}, U) \le \epsilon \in (0, 1/32)$ with probability at least $2/3$, must use at least $\Omega(G/\epsilon)$ queries.

Note that the lower bound holds even for learning algorithms that have a stronger form of access to U than considered for our upper bounds. There, we have only assumed query access to U. In contrast, the lower bound holds even assuming query access to U and $U^\dagger$ as well as controlled versions thereof.

Proof of Proposition 9. The proof builds on the following lemma that maps the problem into a fractional-query one [23].

Lemma 24 (Reduction to fractional-query algorithms [23, Lemma 4.5 and proof of Theorem 1.2]). Let $R \in U(d)$ be a Hermitian unitary (i.e., $R^2 = I$). Define $R^\alpha = (I+R)/2 + e^{-i\pi\alpha}(I-R)/2$ for some $\alpha \in (0, 1]$. Suppose that there exists an algorithm $\mathcal{A}$ that uses Q queries to

R^α or $R^{\alpha\dagger}$ and produces some output with probability at least $2/3$. Then there exists another algorithm $\mathcal{A}'$ that uses $50 + 100\alpha Q$ queries to controlled-R and produces the same output with probability at least $\exp(-\alpha\pi Q)/2$.

To use this lemma, we need to construct a packing net of Hermitian unitaries and give an upper bound on the maximum probability of successfully distinguishing them. Thus we need the following two lemmas.

Lemma 25 (Packing net of Hermitian unitaries; variant of Ref. [23, Proposition 4.1]). There exists a set of Hermitian unitaries $\mathcal{P} = \{R_i\}_i \subset U(d)$ with $\log|P| \geq \Omega(d^2)$ and $R_i^2 = I$ for $R_i \in \mathcal{P}$, such that for any $R_i \neq R_j \in \mathcal{P}$, $d'_F(R_i, R_j) \geq 1/8$.

Proof. Let $d = 2r + 1$ if d is odd, or $d = 2r + 2$ if d is even. In Ref. [145, Lemma 7] (or Ref. [21, Lemma 8]), it is asserted that there exists a set of rank-r density matrices in dimension $2r$ with cardinality at least $\exp\left(r^2/8\right)$, such that all the nonzero eigenvalues are equal to $1/r$ and any two different density matrices have trace distance at least $1/4$. We can write this set as $\{(I_{2r} + V_i)/(2r), i = 1, \ldots, N\}$, where $V_i \in U(2r)$ is a Hermitian unitary of trace zero. Then, $N \geq \exp\left(r^2/8\right)$, and $\forall i \neq j$,

$$\frac{1}{4} \leq \frac{1}{2}\left\|\frac{I_{2r} + V_i}{2r} - \frac{I_{2r} + V_j}{2r}\right\|_1 = \frac{1}{4r}\|V_i - V_j\|_1 \leq \frac{1}{\sqrt{2r}}\|V_i - V_j\|_F, \tag{C37}$$

where we have used $\|V_i\|_1 \leq \sqrt{2r}\|V_i\|_F$. Then, we embed $V_i \to R_i = V_i \oplus I_b \in U(d)$, where $b = 1$ or 2, depending on whether d is odd or even. We have

$$d_F(R_i, R_j) = \frac{1}{\sqrt{d}}\|R_i - R_j\|_F \geq \frac{1}{4}\sqrt{\frac{2r}{2r + b}} \geq \frac{1}{8}. \tag{C38}$$

Now, we would like to translate d_F into d'_F. From Lemma 10, we know that changing to the quotient metric for any set of unitaries only decreases $\log N$ by an additive constant (since here we consider constant ϵ). Therefore, we still have $\log|P| \geq \Omega(d^2)$ for $d'_F(R_i, R_j) \geq \frac{1}{8}$. ■

Lemma 26 (Upper bound on success probability of distinguishing unitaries [68, Theorem 5]). Let $\mathcal{P} \subseteq U(d)$ be a set of unitaries. Let $\mathcal{A}$ be any algorithm that uses Q queries to an input unitary U_x and output a guess $\hat{x}$. Suppose that the input unitary is randomly picked from $\mathcal{P}$ with uniform probability. Then, the maximal probability that the output satisfies $\hat{x} = x$ is upper bounded by $(1/|\mathcal{P}|)\binom{Q+d^2-1}{Q}$.

Now, we can proceed to prove the lower bound in Proposition 9. Suppose that we have a learning algorithm $\mathcal{A}$ that uses Q queries and outputs a $\hat{U}$ that has accuracy ϵ in d_{avg} with success probability at least $2/3$. From the theory of universal gates [146], we know that $G = \mathcal{O}(4^k)$ gates suffice to implement an arbitrary k-qubit unitary, i.e., there exists a constant C such that G gates can implement an arbitrary unitary on $k = \lfloor \log_4(G/C) \rfloor$ qubits. Let $d = \min\{2^n, 2^k\}$, and focus on the first $\min\{n, k\}$ qubits. The algorithm $\mathcal{A}$ is thus able to learn any unitary on these qubits.

Consider the packing net $\mathcal{P} = \{R_i\}$ from Lemma 25 for this choice of d. We want to identify $R \in \mathcal{P}$ but using only access to R^α for $1/\alpha = \lfloor 1/32\epsilon \rfloor > 1$. If we apply $\mathcal{A}$ to R^α, then with probability at least $2/3$, the output U satisfies $d_{\text{avg}}(U, R^\alpha) \leq \epsilon$. From the equivalence of d_{avg} and d'_F (Lemma 4), and the triangle inequality and unitary invariance of d'_F, we have

$$d'_F(U^{1/\alpha}, R) \leq \sum_{p=1}^{1/\alpha} d'_F(U^p R^{1-\alpha p}, U^{p-1} R^{1-\alpha p+\alpha}) = \frac{2}{\alpha} d_{\text{avg}}(U, R^\alpha) \leq \frac{2\epsilon}{\alpha} \leq \frac{1}{16}. \tag{C39}$$

Since $R \in \mathcal{P}$ have pairwise distance at least $1/8$, the algorithm can identify R with success probability at least $2/3$ by finding the closest element of $\mathcal{P}$ to $U^{1/\alpha}$.

Now, via Lemma 24, we know that there is a learning algorithm $\mathcal{A}'$ that can use $50 + 100\alpha Q$ queries to controlled-R to identify R with success probability at least $\exp(-\alpha\pi Q)/2$. On the other hand, we know that the success probability cannot exceed the upper bound $\binom{Q+d^2-1}{Q}/|\mathcal{P}|$ set by Lemma 26 with $\log|P| \geq \Omega(d^2)$. Combined with a technical lemma [23, Lemma 4.3], this means that the number of queries must be at least $\Omega(d^2)$, i.e.,

$$50 + 100\alpha Q \geq \Omega(d^2) \implies Q \geq \Omega\left(\frac{d^2}{\alpha}\right) = \Omega\left(\frac{d^2}{\epsilon}\right) = \Omega\left(\frac{\min\{4^n, G\}}{\epsilon}\right). \tag{C40}$$

This concludes the proof of Proposition 9. ■

We comment on the connection of our results to the recent work [56] on the hardness of learning Haar-random unitaries, where the authors have proved a sample-complexity lower bound $\Omega\left(d^2/\log^2 d\right)$ for learning d-dimensional Haar-random unitaries to constant accuracy with respect to d'_F. The direct consequence of our lower bound when applied to learning the whole unitary group $U(d)$, without assumptions of limited complexity, is a lower bound of $\Omega\left(d^2\right)$, which is stronger than that of Ref. [56, Theorem 1] by a factor of $\log^2 d$. We note that

this difference is a consequence of proof techniques that comes about in two ways. One $\log d$ factor comes from the authors' analysis of the differential entropy, which only calculates the contribution of $\Theta(d/\log d)$ columns of the matrix elements, instead of all d columns. This issue does not arise for us, because we focus on the discrete entropy with the use of a packing net. The other $\log d$ comes from the mutual information upper bound, where the authors use the straightforward Holevo bound: each d-dimensional quantum state can carry at most $\mathcal{O}(\log d)$ bits of information. We manage to get rid of this factor by making use of a more refined bound on the success probability, as in Lemma 26.

Lastly, we remark on the proof technique used here compared to the Holevo information bound in the state-learning case (Appendix B 2). The Holevo bound is particularly useful in proving these lower bounds because, combined with the data-processing inequality, it gives an upper bound on the amount of information that can be extracted from quantum states. In particular, it asserts that, given an ensemble of d-dimensional states $\{\rho_X\}$ with random classical labels $X \in [M]$, the maximal mutual information with the underlying random label when using k copies of the state is upper bounded by $\chi(X;\rho_X^{\otimes k}) \triangleq S(\mathbb{E}_X[\rho_X^{\otimes k}]) - \mathbb{E}_X[S(\rho_X^{\otimes k})]$. Meanwhile, the information needed to distinguishing a packing net of d-dimensional states is lower bounded by $\Omega(d)$. Thus, upper bounding the Holevo χ via the number of samples k can give us sample-complexity lower bounds. A naive upper bound is $\chi \le S(\mathbb{E}_X[\rho_X^{\otimes k}]) \le k\log d$, because $\mathbb{E}_X[\rho_X^{\otimes k}]$ is a d^k-dimensional mixed state and thus has entropy at most $k\log d$. This gives us a $\Omega(d/\log d)$ sample-complexity lower bound with a suboptimal logarithmic factor. To get rid of the $\log d$ factor, in Ref. [179] it has been noted that k copies of a d-dimensional pure state live in the symmetric subspace of the k-fold tensor power of d-dimensional Hilbert space. Therefore, the first term $S(\mathbb{E}_X[\rho_X])$, along with the Holevo χ, can be more tightly upper bounded by $\log\binom{k+d-1}{k}$, where the binomial coefficient is the dimension of the symmetric subspace. This can then be used to prove a $\Omega(d)$ lower bound, which is optimal in d.

However, an analogous result for unitary (or, more generally, channel) queries is still lacking. Consider an ensemble of channels $\{\mathcal{C}_X\}$ labeled by a classical random variable $X \in [M]$. In general, one can sequentially query the channel k times interleaved with processing operations to prepare a state carrying the information extracted from the queries. This then has the form $\rho_X^k = \mathcal{C}_k\mathcal{C}_X\mathcal{C}_{k-1}\mathcal{C}_X\cdots\mathcal{C}_1\mathcal{C}_X(\rho^0)$, where the $\mathcal{C}_i$ are fixed channels independent of X and ρ_0 is some fixed state. Then, the amount of information that one can extract is given by the Holevo information $\chi(X;\rho_X^k) = S(\mathbb{E}_X[\rho_X^k]) - \mathbb{E}_X S([\rho_X^k])$. Upper bounding this quantity is in general difficult. In Ref. [180], the authors have used induction and obtained $\chi \le k\log(d^2)$, which corresponds to the naive upper bound in the state case. Using this, however, can only give us a suboptimal $\Omega(d^2/\log d)$ query-complexity lower bound. We suspect that an improved method, similar in spirit to Ref. [179], making use of the fact that all k queries are to the same channel $\mathcal{C}_X$ should be possible and should give a

$$\chi(X;\rho_X^k) \le \log\binom{k+d^2-1}{k} \tag{C41}$$

upper bound. This would then also give an information-theoretic perspective on the binomial coefficient appearing in the unitary discrimination result of Lemma 26, originally proved by positive-semidefinite programming. We leave the proof of this Holevo information bound as an open problem for future work.

4. Learning from classically described data

As we have seen in Theorems 1 and 4, the sample complexities of learning G-gate states and unitaries are both $\tilde{\Theta}(G)$. This suggests that they have similar sources of complexity. However, differently from state learning, we can identify two sources of difficulty in unitary learning: (1) reading out the input and output quantum states, and (2) learning the mapping from inputs to outputs. The similar complexity $\tilde{\Theta}(G)$ of both state and unitary learning suggests that learning the mapping is actually easy and may only need a constant number of queries to the unknown unitary.

To formalize this idea, we consider a different access model for the unitary-learning task: we focus on learning the mapping by assuming training data that contains classical descriptions of input and output states. Specifically, we consider a learning algorithm H that selects N input n-qubit states $\{|x_i\rangle\}_{i=1}^N$ and queries the unknown unitary to obtain $\{U|x_i\rangle\}_{i=1}^N$, where we have (repeated) access to the classical descriptions of all these input and output states. Based on these classically described data, we want to use the learning algorithm H to output a $\hat{U}$ that satisfies $d_{\mathrm{avg}}(\hat{U},U) \le \epsilon$.

A recent line of research on the quantum no-free-lunch theorem [50,51] implies that the above task of learning the mapping from classically described data in the average-case distance requires at least $\Omega(2^n)$ samples. This seems to contradict our idea that learning the mapping should be easy. However, in Ref. [51], how to circumvent the quantum no-free-lunch theorem has also been demonstrated. In particular, the authors have shown that by entangling our input states with an ancillary system, applying the unitary on the original system, and collecting the output entangled states, we can reduce the sample requirement by a factor equal to the Schimidt rank r of the entangled states. In the limit of the maximally entangled state, where $r = 2^n$, the output state is in fact the Choi-Jamiołkowski state of the unitary, which already contains all the matrix elements of

the unitary. Therefore, in Ref. [51] it has been concluded that the use of entangled data can reduce the data requirements and eventually make the unitary-learning task easy, requiring only one sample with a maximally entangled input state.

Here, we aim to go beyond this result and provide a unified information-theoretic reformulation of the quantum no-free-lunch theorem (Theorem 17), which is not limited to entangled data. We find that the key ingredient to reduce the sample complexity of learning with classical description is to enlarge the representation space (i.e., the space in which the output states live). While entanglement is one way to achieve such an enlargement, it is not the only one. In fact, we find an alternative method that only uses classically mixed states and achieves the same reduction in sample complexity. Specifically, we establish the following theorem.

Proposition 10 (Upper bounds in learning with classical descriptions; restatement of upper bounds in Theorem 5). There exists a learning algorithm H_{entangle} that, for any n-qubit unitary $U \in U(2^n)$, uses $N = \lceil 2^n/r \rceil$ classically described data $\{(|x_i\rangle, (U \otimes I)|x_i\rangle)\}_{i=1}^N$, where the $|x_i\rangle$ are bipartite entangled states over two n-qubit systems with Schmidt rank at most r, to output a $\hat{U}$ such that $d_{\text{avg}}(\hat{U}, U) \leq \epsilon$ for any $\epsilon > 0$.

Similarly, there exists a learning algorithm H_{mixed} that, for any n-qubit unitary $U \in U(2^n)$, uses $N = \lceil 2^n/r \rceil$ classically described data $\{(\rho_i, (U \otimes I)\rho_i(U \otimes I)^\dagger)\}_{i=1}^N$, where the ρ_i are classically mixed states over two n-qubit systems with rank at most r, of the form

$$\rho_i = \sum_{j=1}^{r} p_{ij} \left|\phi_{ij}\right\rangle\!\left\langle\phi_{ij}\right| \otimes \left|\psi_{ij}\right\rangle\!\left\langle\psi_{ij}\right| , \tag{C42}$$

to output a $\hat{U}$ such that $d_{\text{avg}}(\hat{U}, U) \leq \epsilon$ for any $\epsilon > 0$.

Note that, since the number N of training data points in Proposition 10 is independent of the desired accuracy $\epsilon > 0$, we can also learn with respect to $d_\diamond$. In fact, we can even learn the unknown unitary exactly.

We prove Proposition 10 by explicitly constructing the learning algorithms. We remark that the $r = 2^n$ case for entangled data has previously appeared in Ref. [51] and that a different strategy using mixed states has been proposed in Ref. [73].

Proof. Let $d = 2^n$. We begin by describing the algorithm for entangled data. We consider the following set of input states:

$$\left|x_j\right\rangle = \frac{1}{\sqrt{\mathcal{Z}_j}} \sum_{i=(j-1)r+1}^{\min\{jr,d\}} |i\rangle \otimes |i\rangle , \quad j = 1, \ldots, \lceil d/r \rceil , \tag{C43}$$

where the normalization $\mathcal{Z}_j = r$ for $1 \leq j \leq \lceil d/r \rceil - 1$ and $\mathcal{Z}_j = d - (\lceil d/r \rceil - 1)r$ for $j = \lceil d/r \rceil$. They all have Schmidt rank at most r. If we apply $U \otimes I$ on $\left|x_j\right\rangle$, the output state reads

$$(U \otimes I)\left|x_j\right\rangle = \frac{1}{\sqrt{\mathcal{Z}_j}} \sum_{i=(j-1)r+1}^{\min\{jr,d\}} \sum_{k=1}^{d} \langle k|U|i\rangle\, |k\rangle \otimes |i\rangle . \tag{C44}$$

Since we have the classical description, we can directly read off the matrix elements $\langle k|U|i\rangle$ with $1 \leq k \leq d$ and $(j-1)r + 1 \leq i \leq \min\{jr, d\}$. Combining different j, we can gather all the matrix elements that we need to learn U.

Next, we describe the algorithm for mixed-state data. We consider the input states to be

$$\rho_j = \sum_{i=(j-1)r+1}^{\min\{jr,d\}} p_j\, |ii\rangle\langle ii| , \quad j = 1, \ldots, \lceil d/r \rceil . \tag{C45}$$

where the uniform mixing probability $p_j = 1/r$ for $1 \leq j \leq \lceil d/r \rceil - 1$ and $p_j = 1/(d - (\lceil d/r \rceil - 1)r)$ for $j = \lceil d/r \rceil$. Then, all ρ_j have rank at most r. If we apply $U \otimes I$ on $\left|x_j\right\rangle$, the output state becomes

$$\rho_j = \sum_{i=(j-1)r+1}^{\min\{jr,d\}} p_j\, (U \otimes I)\, |ii\rangle\langle ii|\, (U \otimes I)^\dagger, \quad j = 1, \ldots, \lceil d/r \rceil . \tag{C46}$$

We can interpret this output mixed state as randomly choosing a basis state in the ancillary system and applying the unitary to the same state in the original system. Since we have the classical description, we can use the ancillary system as a label for the state that we have inputted (e.g, $|i\rangle$) and read off all the amplitudes of $U|i\rangle$ on the original system, i.e., a column of the U matrix. Then, by combining all the different basis elements $|i\rangle$, $1 \leq i \leq d$, we obtain all the matrix elements of U. ■

Now, we move on to the lower bound, which states that any noise-robust unitary-learning algorithm needs at least $\Omega(2^n/r)$ samples to learn an arbitrary unknown unitary from classically described data. The noise-robust requirement here is in accordance with realistic learning scenarios in which the tomography of input and output states necessarily involves reconstruction imperfection and noise. Specifically, we have the following proposition.

Proposition 11 (Lower bounds in learning with classical descriptions; restatement of lower bounds in Theorem 5). Let $\epsilon \in (0, 1), \eta = \Theta(\epsilon)$. Let H_{entangle} be any learning algorithm that, for any n-qubit unitary $U \in U(2^n)$, uses classically described data $\{(|x_i\rangle, |y_i\rangle)\}_{i=1}^N$, where the $|x_i\rangle$

are bipartite entangled states over two n-qubit systems with Schmidt rank at most r and the $|y_i\rangle$ are η-noisy versions of $(U\otimes I)\,|x_i\rangle$ satisfying $d_{\mathrm{tr}}(|y_i\rangle\,,(U\otimes I)\,|x_i\rangle)\le\eta$, to output a $\hat{U}$ such that $d_{\mathrm{avg}}(\hat{U},U)\le\epsilon$. Then, H_{entangle} needs at least $N\ge\Omega(2^n/r)$ samples.

Similarly, let H_{mixed} be any learning algorithm that, for any n-qubit unitary $U\in U(2^n)$, uses classically described data $\{(\rho_i,\sigma_i)\}_{i=1}^N$, where the ρ_i are classically mixed states over two n-qubit systems with rank at most r of the form

$$\rho_i=\sum_{j=1}^{r}p_{ij}\,|\phi_{ij}\rangle\langle\phi_{ij}|\otimes|\psi_{ij}\rangle\langle\psi_{ij}| \tag{C47}$$

and the σ_i are η-noisy versions of $(U\otimes I)\rho_i(U\otimes I)^\dagger$ satisfying $d_{\mathrm{tr}}(\sigma_i,(U\otimes I)\rho_i(U\otimes I)^\dagger))\le\eta$, to output a $\hat{U}$ such that $d_{\mathrm{avg}}(\hat{U},U)\le\epsilon$. Then, H_{mixed} needs at least $N\ge\Omega(2^n/r)$ samples.

Proposition 11 is a consequence of the following information-theoretic reformulation of the quantum no-free-lunch theorem. The intuition behind this theorem is simple. On the one hand, to learn a unitary, we have to gather enough information to specify it. This required amount of information is quantified by the metric entropy of the unitary class. On the other hand, the information provided by each sample is limited and can be characterized by the metric entropy of the output state space. Therefore, the number of samples needed to learn the unitary is given by the former divided by the latter. In particular, we can see that the data requirement can be reduced if we increase the amount of information carried by each sample, represented by the metric entropy in the denominator.

Theorem 17 (Information-theoretic reformulation of quantum no-free-lunch theorem). Let $\eta,\epsilon\in(0,1)$. Let S be a set of input states (possibly with ancillas) and let P be a distribution over S. Let $\{\rho_i\}_{i=1}^N\subset S^N$ be N classically described input states. Suppose that after applying the unknown n-qubit unitary U from a class $\mathcal{U}\subseteq U(2^n)$ of unitaries, they are transformed into the output states $\{\sigma_i\}_{i=1}^N$ through the map $f_U:\rho_i\mapsto\sigma_i=f_U(\rho_i)$. Let $\tilde{\sigma}_i$ be an η-noisy version of σ_i satisfying $d_{\mathrm{tr}}(\tilde{\sigma}_i,\sigma_i)\le\eta$. Let $\mathcal{N}_\eta=\sup_{\rho\in S}\mathcal{N}(\{f_V(\rho):V\in\mathcal{U}\},d_{\mathrm{tr}},\eta)$ be the maximal covering number of the set of all possible output states with different unitaries acting on the input states. Let $\mathcal{F}_{\mathcal{U}}=\{f_V:V\in\mathcal{U}\}$ be the set of maps and let $d_P(f_V,f_W)=\sqrt{\mathbb{E}_{\rho\sim P}[d_{\mathrm{tr}}(f_V(\rho),f_W(\rho))^2]}$ be the root-mean-square trace distance. Then, any learning algorithm H that uses the η-noisy classically described data $\{\rho_i,\tilde{\sigma}_i\}_{i=1}^N$ and outputs a $\hat{U}$ such that $d_P(f_{\hat{U}},f_U)\le\epsilon$ with probability at least $2/3$ needs at least

$$N\ge\Omega\left(\frac{\log\mathcal{M}(\mathcal{F}_{\mathcal{U}},d_P,2\epsilon+6\eta)}{\log\mathcal{N}_\eta}\right) \tag{C48}$$

samples.

In particular, if $\eta=\Theta(\epsilon)$, $\mathcal{U}=U(2^n)$, P is a locally scrambled ensemble up to the second moment over n-qubit pure states (e.g., an n-qubit Haar measure), S is the support of P, and $f_U(\rho)=U\rho U^\dagger$, then at least $\Omega(2^n)$ samples are needed.

We remark that $\eta=\Theta(\epsilon)$ is a convenient choice of noise level for stating the results, but in fact a weaker assumption, $\log(1/\eta)=\Theta(\log(1/\epsilon))$, suffices.

In the following, we will first show that Theorem 17 implies Proposition 11 (the lower bounds in Theorem 5). Then, we will turn to the proof of Theorem 17.

Proof of Proposition 11. In both cases (entangled or mixed), we prove the $\Omega(2^n/r)$ lower bound in two steps via Theorem 17: (1) we show that the numerator $\log\mathcal{M}(\mathcal{F}_{\mathcal{U}},d_P,2\epsilon+6\eta)$ in Theorem 17 is at least $\Omega(4^n\log(1/\epsilon))$; and (2) we show that the denominator $\mathcal{N}_\epsilon$ is at most $\mathcal{O}(2^n r\log(1/\epsilon))$ when the input states are either entangled pure states of Schmidt rank at most r or mixed states of rank at most r. Then, the desired results follow.

For step (1), we begin by defining the distribution P with respect to which the performance in Theorem 17 is measured. For both entangled and mixed cases, we define P to be the distribution of $|\psi\rangle\otimes|0\rangle^{\otimes n}$, where $|\psi\rangle$ is a Haar-random state on the original system and $|0\rangle^{\otimes n}$ is a fixed state on the ancillary system. Note that this state is indeed both a bipartite entangled state with Schmidt rank at most r and of the form given in Eq. (C47), with rank at most r. Moreover, since, in both cases, the map f_U is given by acting the unitary U on the original system and the identity on the ancillary system, the distance metric $d_P(f_V,f_W)$ is the same as d_{avg}. Therefore, the packing number satisfies

$$\mathcal{M}(\mathcal{F}_{U(2^n)},d_P,2\epsilon+6\eta)=\mathcal{M}(U(2^n),d_{\mathrm{avg}},2\epsilon+6\eta). \tag{C49}$$

To find the packing number $\mathcal{M}(U(2^n),d_{\mathrm{avg}},2\epsilon+6\eta)$, we invoke the covering-number bound for $U(2^n)$ with respect to the normalized Frobenius norm d_F (Lemma 9), the fact that quotienting out the global phase only changes the metric entropy by a constant (Lemma 10) and the equivalence of d_F' and d_{avg} [Lemma 4, item (1)]. We have

$$\begin{aligned}&\log\mathcal{M}(\mathcal{F}_{U(2^n)},d_P,2\epsilon+6\eta)\\&\quad=\log\mathcal{M}(U(2^n),d_{\mathrm{avg}},2\epsilon+6\eta)\ge\Omega\left(4^n\log\frac{1}{\epsilon}\right),\end{aligned} \tag{C50}$$

where we have used $\eta=\Theta(\epsilon)$.

Next, for step (2), we compute $\mathcal{N}_\eta$. For entangled data, note that applying unitaries on only the first n qubits does not change the bipartite Schmidt rank r, so the output states are pure states of the form $|\chi\rangle=\sum_{i,j=1}^{2^n}A_{ij}\,|i\rangle\otimes|j\rangle$,

where $\|A\|_F = 1$ because of normalization, and the rank of A corresponds to the Schmidt rank, which is at most r. Furthermore, the Euclidean distance between the output states is equal to the Frobenius distance between the corresponding A-matrices. With this correspondence, we can explicitly construct a covering net over the output states as follows. We take a minimal η-covering net $\mathcal{N}'$ over the set of complex matrices A with bounded rank r and $\|A\|_F = 1$ with respect to the Frobenius distance. Since they are contained in the unit ball ($\|A\|_F \leq 1$) in a real linear space of dimension $2 \cdot 2^n \cdot r$ [181, Theorem 1], by the monotonicity of the covering number and the standard covering-number bound for Euclidean balls via a volume argument [53, Corollary 4.2.13], we have $\log|\mathcal{N}'| \leq \mathcal{O}(2^n r \log(1/\eta))$. Meanwhile, similarly to the proof in Lemma 1, the trace distance between any two pure states $|\psi\rangle, |\phi\rangle$ is bounded by the Euclidean distance and thus the Frobenius distance between the corresponding A matrices is as follows:

$$\begin{aligned} d_{\mathrm{tr}}(|\psi\rangle, |\phi\rangle) &= \sqrt{1 - |\langle\psi|\phi\rangle|^2} \leq \sqrt{2(1 - |\langle\psi|\phi\rangle|)} \\ &\leq \sqrt{2(1 - \mathrm{Re}[\langle\psi|\phi\rangle])} = \||\psi\rangle - |\phi\rangle\|_2. \end{aligned} \tag{C51}$$

Therefore, $\mathcal{N}'$ gives an η-covering net over the output states with respect to the trace distance d_{tr}. Hence, $\log\mathcal{N}_\eta \leq \log|\mathcal{N}'| \leq \mathcal{O}(2^n r \log(1/\eta)) = \mathcal{O}(2^n r \log(1/\epsilon))$, since $\eta = \Theta(\epsilon)$, and from Theorem 17 we have the desired lower bound,

$$N \geq \Omega\left(\frac{4^n \log(1/\epsilon)}{2^n r \log(1/\epsilon)}\right) = \Omega\left(\frac{2^n}{r}\right). \tag{C52}$$

The case of mixed states is similar. For a given input state $\rho = \sum_{i=1}^r p_i |\phi_i\rangle\langle\phi_i| \otimes |\psi_i\rangle\langle\psi_i|$, the output state reads

$$\sigma = \sum_{i=1}^r p_i U|\phi_i\rangle\langle\phi_i| U^\dagger \otimes |\psi_i\rangle\langle\psi_i|. \tag{C53}$$

Now, we take a minimal η-covering net $\mathcal{N}''$ over all pure n-qubit states with respect to the Euclidean distance, which is a unit ball in a $(2 \cdot 2^n)$-dimensional real linear space. By the standard covering-number bound for Euclidean balls, we know that $\log|\mathcal{N}''| \leq \mathcal{O}(2^n \log(1/\eta)) = \mathcal{O}(2^n \log(1/\epsilon))$. Then, for any $U|\phi_i\rangle$, there exists a $|\eta_i\rangle \in \mathcal{N}''$ such that $\|U|\phi_1\rangle - |\eta_1\rangle\|_2 \leq \eta$. Let $\sigma' = \sum_{i=1}^r p_i |\eta_i\rangle\langle\eta_i| \otimes |\psi_i\rangle\langle\psi_i|$. Then, the trace distance is bounded by

$$\begin{aligned} \frac{1}{2}\|\sigma - \sigma'\|_1 &\leq \frac{1}{2}\sum_{i=1}^r p_i \|U|\phi_i\rangle\langle\phi_i| U^\dagger \otimes |\psi_i\rangle\langle\psi_i| - |\eta_i\rangle\langle\eta_i| \otimes |\psi_i\rangle\langle\psi_i|\|_1 \\ &\leq \frac{1}{2}\sum_{i=1}^r p_i \|U|\phi_i\rangle \otimes |\psi_i\rangle - |\eta_i\rangle \otimes |\psi_i\rangle\|_2 \\ &= \frac{1}{2}\sum_{i=1}^r p_i \|U|\phi_i\rangle - |\eta_i\rangle\|_2 \leq \frac{\eta}{2}\sum_{i=1}^r p_i = \frac{\eta}{2}, \end{aligned} \tag{C54}$$

where we have used the subadditivity of the trace norm, the fact that the trace distance is upper bounded by the Euclidean norm for pure states, and $\sum_{i=1}^r p_i = 1$. Hence the set

$$\left\{\sum_{i=1}^r p_i |\eta_i\rangle\langle\eta_i| \otimes |\psi_i\rangle\langle\psi_i| : |\eta_i\rangle \in \mathcal{N}'', 1 \leq i \leq r\right\} \tag{C55}$$

forms an $\eta/2$-covering net of set of the output states and has cardinality $|\mathcal{N}''|^r$. Therefore, we have $\log\mathcal{N}_\eta = r\log|\mathcal{N}''| = \mathcal{O}(2^n r \log(1/\epsilon))$. From Theorem 17, we again arrive at the desired result,

$$N \geq \Omega\left(\frac{4^n \log(1/\epsilon)}{2^n r \log(1/\epsilon)}\right) = \Omega\left(\frac{2^n}{r}\right). \tag{C56}$$

This concludes the proof of Proposition 11 and together with Proposition 10, we have proved Theorem 5. ■

Now, we move on to prove our quantum no-free-lunch theorem (Theorem 17). We first establish the following information-theoretic lower bound on the sample complexity of learning discrete functions. We remark that a version for binary-valued functions has been proved in a different

fashion in Ref. [182, Proposition 8] and Ref. [183, Lemma 4.8].

Proposition 12 (Information-theoretic lower bound for learning discrete functions). Let $\epsilon > 0$, $k \in \mathbb{N}$, and $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathcal{Y} = \{1, \ldots, k\}$ with a distance metric d. Any learning algorithm H that uses N samples $\{x_i \in \mathcal{X}, y_i = f(x_i)\}_{i=1}^N$ and outputs an $\hat{f}$ such that $d(\hat{f}, f) \leq \epsilon$ with probability at least 2/3 for any $f \in \mathcal{F}$ must use at least

$$N \geq \Omega\left(\frac{\log \mathcal{M}(\mathcal{F}, d, 2\epsilon)}{\log k}\right) \tag{C57}$$

samples.

Proof of Proposition 12. We begin by taking a maximal 2ϵ-packing $\mathcal{P}$ of $\mathcal{F}$, i.e., $|\mathcal{P}| = \mathcal{M}(\mathcal{F}, d, 2\epsilon)$ and for any $f_i \neq f_j \in \mathcal{P}$, $d(f_i, f_j) > 2\epsilon$. Now, we design a communication protocol between two parties, Alice and Bob, as follows. The packing $\mathcal{P}$ is shared by both parties. Alice takes a random variable W uniformly sampled from $\{1, \ldots, \mathcal{M}(\mathcal{F}, d, 2\epsilon)\}$ and picks the corresponding function f_W from the packing $\mathcal{P}$. She then feeds the inputs $x = (x_1, \ldots, x_N)$ into f_W, generating a data set $Z = ((x_1, f_W(x_1)), \ldots, (x_N, f_W(x_N)))$, and sends the data set to Bob. Bob's task is to use this data set to determine which function Alice has used. Suppose that Bob is given a learning algorithm described as in the proposition. The algorithm will learn from the data set and output a hypothesis function $\hat{f}$ that satisfies

$$\mathbb{P}[d(\hat{f}, f_W) \leq \epsilon] \geq 2/3, \tag{C58}$$

no matter which W was chosen by Alice. With $\hat{f}$ in hand, Bob can make the guess

$$\hat{W} = \operatorname{argmin}_{f_w \in \mathcal{P}} d(\hat{f}, f_w). \tag{C59}$$

Note that as long as $d(\hat{f}, f_W) \leq \epsilon$, then, for any $f_i \neq f_W \in \mathcal{P}$,

$$d(\hat{f}, f_i) \geq d(f_W, f_i) - d(\hat{f}, f_W) > 2\epsilon - \epsilon = \epsilon \geq d(\hat{f}, f_W). \tag{C60}$$

Therefore, the error probability of Bob's guess is bounded by

$$\begin{aligned}\mathbb{P}[\hat{W} \neq W] &= \mathbb{P}[\exists i \neq W : d(\hat{f}, f_i) \leq d(\hat{f}, f_W)] \\ &\leq \mathbb{P}[d(\hat{f}, f_W) > \epsilon] \leq 1/3.\end{aligned} \tag{C61}$$

By the Fano inequality [184, Theorem 2.10.1], the conditional entropy $S(W|\hat{W}) \leq s_2(1/3) + \frac{1}{3}\log \mathcal{M}(\mathcal{F}, d, 2\epsilon)$, where $s_2(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta)$ is the binary entropy function. Then, the mutual information is at least

$$\begin{aligned}I(W; \hat{W}) &= S(W) - S(W|\hat{W}) \\ &\geq (2/3)\log \mathcal{M}(\mathcal{F}, d, 2\epsilon) - s_2(1/3).\end{aligned} \tag{C62}$$

On the other hand, the sample size N controls the amount of information to which Bob has access. Since Bob's guess is produced by the data set Z, by the data-processing inequality [184, Theorem 2.8.1], we have

$$\begin{aligned}I(W; \hat{W}) &\leq I(W; Z) = S(Z) - S(Z|W) \\ &= S(f_W(x_1), \ldots, f_W(x_N)) \leq N \log k,\end{aligned} \tag{C63}$$

where we have used $S(Z|W) = 0$, since Z is determined by W, and the fact that $(f_W(x_1), \ldots, f_W(x_N))$ can take no more than k^N different values. Combining the above two inequalities, we arrive at

$$N \geq \Omega\left(\frac{\mathcal{M}(\mathcal{F}, d, 2\epsilon)}{\log k}\right). \tag{C64}$$

■

With Proposition 12, we can prove Theorem 17 by quantizing the output states to the nearest elements in covering nets, similar to an idea employed in Ref. [185].

Proof of Theorem 17. Let $k = \mathcal{N}_\eta$. Since $\mathcal{N}_\eta \geq \mathcal{N}(\{f_V(\rho), V \in \mathcal{U}\}, d_{\mathrm{tr}}, \eta)$ for every $\rho \in S$, we can find an η-covering net $\mathcal{N}_\rho$ of size k for each $\rho \in S$. We label the elements of $\mathcal{N}_\rho$ using $\{1, \ldots, k\}$ and define $L_\rho(\sigma) \in [k]$ as the label of a covering-net element $\sigma \in \mathcal{N}_\rho$.

Now, we define the quantized function Qf_U that maps an input state ρ to an element of the covering net $\mathcal{N}_\rho$. Specifically, for any $\rho \in S$ and any $\sigma \in \{f_V(\rho), V \in \mathcal{U}\}$, there exists a $\sigma' \in \mathcal{N}_\rho$, such that $d_{\mathrm{tr}}(\sigma, \sigma') \leq \eta$. For any unitary $U \in \mathcal{U}$, we define

$$Qf_U(\rho) = \operatorname{argmin}_{\sigma \in \mathcal{N}_\rho} d_{\mathrm{tr}}(f_U(\rho)\sigma) \tag{C65}$$

and let $LQf_U(\rho) = L_\rho[Qf_U(\rho)]$ be the corresponding label. (Ties are broken arbitrarily.) Then, LQf_U is a discrete-output function mapping input states S to labels $[k]$ and it is in one-to-one correspondence with Qf_U. We use $\mathcal{F}^Q$ to denote all these labeled quantized functions, $\mathcal{F}^Q = \{LQf_U, U \in \mathcal{U}\}$, and define the distance metric on labeled functions as $d_L(LQf_V, LQf_W) = d_P(Qf_V, Qf_W)$. A useful property is that for any unitary $U \in \mathcal{U}$, we have

$$\begin{aligned}d_P(f_U, Qf_U) &= \sqrt{\mathbb{E}_{\rho \sim P}[d_{\mathrm{tr}}(f_U(\rho), Qf_U(\rho))^2]} \\ &\leq \sqrt{\mathbb{E}_{\rho \sim P}[\eta^2]} = \eta.\end{aligned} \tag{C66}$$

Now, we claim that if there exists a noise-robust learning algorithm H for $\mathcal{U}$ to accuracy ϵ in d_P with probability

at least 2/3, then we can use it to construct a learning algorithm $H^{\mathcal{Q}}$ for $\mathcal{F}^{\mathcal{Q}}$ to accuracy $\epsilon + 2\eta$ in d_L with success probability at least 2/3. Hence, the sample complexity for $\mathcal{U}$ must satisfy

$$N \geq \Omega\left(\frac{\log \mathcal{M}(\mathcal{F}^{\mathcal{Q}}, d_L, 2\epsilon + 4\eta)}{\log k}\right), \tag{C67}$$

by Proposition 12.

To show this claim, we construct $H^{\mathcal{Q}}$ as follows. For any $LQf_U \in \mathcal{F}^{\mathcal{Q}}$, let the data set be

$$Z = (\rho_1, Qf_U(\rho_1)), \ldots, (\rho_N, Qf_U(\rho_N)). \tag{C68}$$

From the definition of quantized functions, we know that the $Qf_U(\rho_i)$ are the η-noisy version of $f_U(\rho_i)$ because $d_{\mathrm{tr}}(Qf_U(\rho_i), f_U(\rho_i)) \leq \eta$. Now, we define $H^{\mathcal{Q}}$ as

$$H^{\mathcal{Q}}[Z] = LQf_{H[Z]}. \tag{C69}$$

Since the learning algorithm H is η-noise-robust, we have $d_P(f_{H[Z]}, f_U) \leq \epsilon$ and thus $d_P(f_{H[Z]}, Qf_U) \leq d_P(f_{H[Z]}, f_U) + d_P(f_U, Qf_U) \leq \epsilon + \eta$ with probability at least 2/3. Then, by the triangle inequality (proved similarly as in Lemma 6), we have

$$\begin{aligned} d_L(H^{\mathcal{Q}}[Z], LQf_U) = d_P(Qf_{H[Z]}, Qf_U) &\leq d_P(Qf_{H[Z]}, f_{H[Z]}) \\ &+ d_P(f_{H[Z]}, Qf_U) \leq \epsilon + 2\eta \end{aligned} \tag{C70}$$

with probability at least 2/3. Thus the claim is proved.

At this point, it remains to prove that

$$\mathcal{M}(\mathcal{F}^{\mathcal{Q}}, d_L, 2\epsilon + 4\eta) \geq \mathcal{M}(\mathcal{F}_{\mathcal{U}}, d_P, 2\epsilon + 6\eta). \tag{C71}$$

To prove this, we can take a maximal $(2\epsilon + 6\eta)$-packing $\mathcal{P}$ of $\mathcal{F}_{\mathcal{U}}$ with respect to d_P, with $|\mathcal{P}| = \mathcal{M}(\mathcal{F}_{\mathcal{U}}, d_P, 2\epsilon + 6\eta)$. Then, $\forall f_{U_1} \neq f_{U_2} \in \mathcal{P}$, we have

$$\begin{aligned} 2\epsilon + 6\eta < d_P(f_{U_1}, f_{U_2}) &\leq d_P(f_{U_1}, Qf_{U_1}) + d_P(Qf_{U_1}, Qf_{U_2}) \\ &+ d_P(Qf_{U_2}, U_2) \leq 2\eta + d_P(Qf_{U_1}, Qf_{U_2}). \end{aligned} \tag{C72}$$

Therefore, $d_L(LQf_{U_1}, LQf_{U_2}) = d_P(Qf_{U_1}, Qf_{U_2}) > 2\epsilon + 4\eta$. Hence,

$$\begin{aligned} \mathcal{M}(\mathcal{F}^{\mathcal{Q}}, d_L, 2\epsilon + 4\eta) &\geq |\{LQf_U, U \in \mathcal{P}\}| = |\mathcal{P}| \\ &= \mathcal{M}(\mathcal{F}_{\mathcal{U}}, d_P, 2\epsilon + 6\eta). \end{aligned} \tag{C73}$$

This concludes the proof of the main part of Theorem 17.

Finally, we illustrate the special case in which $\eta = \Theta(\epsilon), \mathcal{U} = U(2^n)$ is the whole unitary group, P is a locally scrambled ensemble up to the second moment over n-qubit pure states (e.g., an n-qubit Haar measure; see Definition 1), S is the support of P, and $f_U(\rho) = U\rho U^\dagger$. We show that at least $\Omega(2^n)$ samples are needed, thus reproducing the quantum no-free-lunch theorem in the usual sense and generalizing it to locally scrambled ensembles.

To see this, we first compute $\log \mathcal{M}(\mathcal{F}_{U(2^n)}, d_P, 2\epsilon + 6\eta)$. From the covering-number bound for $U(2^n)$ with respect to the normalized Frobeinus norm d_F (Lemma 9), the fact that quotienting out the global phase only changes the metric entropy by an additive $\mathcal{O}(\log(1/(2\epsilon + 6\eta)))$ term (Lemma 10), and by the equivalence of d_F', d_{avg}, and d_P [Lemma 4, item (1) and Lemma 5], we know that $\log \mathcal{M}(\mathcal{F}_{U(2^n)}, d_P, 2\epsilon + 6\eta) \geq \Omega\left(4^n \log(1/\epsilon)\right)$, where we have used $\eta = \Theta(\epsilon)$.

Next, we move on to $\mathcal{N}_\eta$. Since the output states are still n-qubit pure states, $\mathcal{N}_\eta$ is the covering number of the set of pure states with respect to d_{tr}. Considering that $\frac{1}{2}\||\psi\rangle\langle\psi|\|_1$ is less than one for any pure state $|\psi\rangle$, the covering number is upper bounded by the covering number of a unit Euclidean ball in a $\Theta(2^n)$-dimensional linear space. Therefore, we have $\log \mathcal{N}_\eta \leq \mathcal{O}(2^n \log(1/\eta)) = \mathcal{O}(2^n \log(1/\epsilon))$ since $\eta = \Theta(\epsilon)$. Hence we arrive at

$$N \geq \Omega\left(\frac{4^n \log(1/\epsilon)}{2^n \log(1/\epsilon)}\right) = \Omega(2^n). \tag{C74}$$

This concludes the proof of Theorem 17. ■

The information-theoretic version of the quantum no-free-lunch theorem (Theorem 17) also gives us a way to generalize the quantum no-free-lunch to a restricted unitary class. For example, for unitaries with bounded circuit complexity G, the packing number in the enumerator is lower bounded by $\Omega(G)$, while the covering number in the denominator is upper bounded by $\mathcal{O}(\min\{G \log G + G \log n, 2^n\})$. This gives us a quantum no-free-lunch theorem for G-gate unitaries, where the sample complexity is lower bounded by $\Omega(1)$ for $G \leq \mathcal{O}(2^n)$, by $\Omega(G/2^n)$ for $\Omega(2^n) < G \leq \mathcal{O}(4^n)$ and $\Omega(2^n)$ for $G \geq \Omega(4^n)$.

5. Computational complexity

Similarly to the state-learning case, our algorithm for average-case unitary learning described in Appendix C 2

is not computationally efficient. In this section, we follow Appendix B 3 and first show that there is no polynomial-time algorithm for learning unitaries composed of $G = \mathcal{O}(n\text{polylog}(n))$ two-qubit gates, assuming that RingLWE cannot be solved efficiently on a quantum computer. This result also holds for unitaries with circuit depth $\mathcal{O}(\text{polylog}(n))$. Then, we invoke a stronger assumption, that RingLWE cannot be solved by any subexponential-time quantum algorithm, and show that any quantum algorithm for learning unitaries composed of $\tilde{\mathcal{O}}(G)$ gates must use $\exp(\Omega(G))$ time. Finally, we explicitly construct an efficient learning algorithm for $G = \mathcal{O}(\log n)$, thus establishing $\log n$ gate complexity as a transition point of computational efficiency.

Theorem 18 (Unitary-learning computational-complexity lower bound assuming polynomial hardness of RingLWE). Let $\lambda = n$ be the security parameter. Let U be a unitary consisting of $G = \mathcal{O}(n\text{polylog}(n))$ gates (or a depth $d = \mathcal{O}(\text{polylog}(n))$ circuit) that implements a pseudorandom function in $\mathcal{RF}$. Such a unitary U exists by Corollary 3. There exists no polynomial-time quantum algorithm for learning a circuit description of U to within $\epsilon \leq 1/64$ average-case distance d_{avg} with probability at least $2/3$ from $N = \text{poly}(\lambda)$ queries, if quantum computers cannot solve RingLWE in polynomial time.

Proof. Suppose, for the sake of contradiction, that there is an efficient algorithm $\mathcal{A}_0$ that can learn a description of U to within ϵ average-case distance with probability at least $2/3$. Then, by standard boosting of success probability (see, e.g., Ref. [23, Proposition 2.4]), there is an efficient algorithm $\mathcal{A}$ that can learn U to the same accuracy with probability at least $p = 1 - 1/8192$ with only a constant-factor overhead in time complexity. Note that this boosting requires the distance metric to be efficiently computable, which is guaranteed by the SWAP test elaborated below. We will construct a polynomial-time quantum distinguisher $\mathcal{D}$ that invokes $\mathcal{A}$ to distinguish between U and the unitary $V \in \mathcal{U}$ corresponding to a random classical function. This contradicts Theorem 11, item (2).

The distinguisher $\mathcal{D}$ operates according to Algorithm 6.

Recall that the SWAP test [150,151] takes two quantum states $|\alpha\rangle, |\beta\rangle$ as input and outputs one with probability $(1 + |\langle\alpha|\beta\rangle|^2)/2$. We denote this algorithm as $\text{SWAP}(|\alpha\rangle, |\beta\rangle)$.

Note that step 2 in Algorithm 6, the preparation of tensor product of one-qubit stabilizer states $|x\rangle, x \in \mathbb{Z}_6^n$, is computationally efficient, because it can be achieved by random one-qubit gates acting on each of the n qubits. Moreover, step 4 can be implemented efficiently on a quantum computer because $\hat{U}$ is given in terms of efficient circuit description and because the SWAP test is efficiently implementable. Thus, assuming the hypothetical learner $\mathcal{A}$ to be efficient, the distinguisher $\mathcal{D}$ is efficient as well.

We analyze the probability that the distinguisher $\mathcal{D}$ outputs 1 when given the pseudorandom function U versus the random classical Boolean function V. We denote the distribution of $|x\rangle$ by Q. From Lemma 5, we have

$$d_Q(U, \hat{U}) = \sqrt{\mathbb{E}_{|x\rangle\sim Q}[d_{\text{tr}}(U|x\rangle, \hat{U}|x\rangle)^2]} \leq \sqrt{2} d_{\text{avg}}(U, \hat{U}). \tag{C75}$$

Case 1: $U \in \mathcal{RF}$. By the guarantees of $\mathcal{A}$, with probability at least p, we have $d_{\text{avg}}(\hat{U}, U) \leq \epsilon \leq 1/64$, where $\hat{U}$ is the unitary learned by algorithm $\mathcal{A}$. This implies that

$$\mathbb{E}_{|x\rangle\sim Q} |\langle x|\hat{U}^\dagger U|x\rangle|^2 = 1 - d_Q^2(U, \hat{U}) \geq 1 - 2\epsilon^2, \tag{C76}$$

where we have used the relationship between the fidelity and the trace distance. Then, it immediately follows from Eq. (C76) that

$$\begin{aligned}
\Pr_{U\in\mathcal{RF},\mathcal{D}}\left[\mathcal{D}^{|U\rangle}(\cdot) = 1\right] &= \Pr_{\substack{U\in\mathcal{RF},|x\rangle\sim Q \\ \mathcal{A},\ \text{SWAP}}}\left[\text{SWAP}\left(U|x\rangle, \hat{U}|x\rangle\right) = 1\right] \\
&= \mathbb{E}_{U\in\mathcal{RF},|x\rangle\sim Q}\left[\Pr_{\mathcal{A},\ \text{SWAP}}\left[\text{SWAP}\left(U|x\rangle, \hat{U}|x\rangle\right) = 1 \,\middle|\, U, |x\rangle\right]\right] \\
&\geq p\,\mathbb{E}_{U\in\mathcal{RF}}\left[\frac{1}{2} + \frac{1}{2}\mathbb{E}_{\hat{U},|x\rangle\sim Q}\left[|\langle x|\hat{U}^\dagger U|x\rangle|^2\right]\right] \\
&\geq p\,\mathbb{E}_{U\in\mathcal{RF}}\left[\frac{1}{2} + \frac{1}{2}(1 - 2\epsilon^2)\right] = p(1 - \epsilon^2) > \frac{8189}{8192},
\end{aligned} \tag{C77}$$

where in the first inequality we split the probability into two terms conditioned on the success and failure of $\mathcal{A}$ and we lower bound the failure term by zero, and in the last inequality we have used the fact that $p(1 - \epsilon^2) \geq (1 - 1/8192)(1 - 1/4096) > 8189/8192$.

Case 2: $U = V \in \mathcal{U}$, where V is the n-qubit unitary implementing a randomly chosen classical function. We want to upper bound the probability that the distinguisher $\mathcal{D}$ outputs 1 when given queries to V. Let $\mathcal{C}$ be the set of all possible output unitaries of $\mathcal{A}$. We follow the same reasoning as in Eq. (C77) and note that

$$\begin{aligned}\Pr_{V\in\mathcal{U},\mathcal{D}}\left[\mathcal{D}^{|V\rangle}(\cdot)=1\right] &\le \mathbb{E}_{V\in\mathcal{U}}\left[\max_{W\in\mathcal{C}}\mathbb{E}_{|x\rangle\sim Q}\left[\frac{1}{2}+\frac{1}{2}|\langle x|V^\dagger W|x\rangle|^2\right]\right]+(1-p)\\ &\le \mathbb{E}_{V\in\mathcal{U}}\left[\max_{W\in\mathcal{C}}\left[1-\frac{1}{4}d_{\mathrm{avg}}(V,W)^2\right]\right]+(1-p)\\ &\triangleq \mathbb{E}_{V\in\mathcal{U}}[O_V]+(1-p),\end{aligned} \tag{C78}$$

where we define $O_V = \max_{W\in\mathcal{C}}\left[1-\frac{1}{4}d_{\mathrm{avg}}(V,W)^2\right]$. Furthermore, we can split the right-hand side into two parts by introducing a constant θ:

$$\begin{aligned}\mathbb{E}_{V\in\mathcal{U}}[O_V] &\le \Pr\left[O_V \le 1-\frac{\theta^2}{4}\right]\cdot\left(1-\frac{\theta^2}{4}\right)\\ &\quad+\Pr\left[O_V > 1-\frac{\theta^2}{4}\right]\cdot 1 \le 1-\frac{\theta^2}{4}\\ &\quad+\Pr\left[O_V > 1-\frac{\theta^2}{4}\right],\end{aligned} \tag{C79}$$

where we have used the fact that $O_V \le 1$. Note that

$$\begin{aligned}\Pr\left[O_V > 1-\frac{\theta^2}{4}\right] &\le \Pr_{V\in\mathcal{U}}\left[\exists W\in\mathcal{C}: d_{\mathrm{avg}}(V,W)<\theta\right]\\ &\le \sum_{W\in\mathcal{N}}\Pr_{V\in\mathcal{U}}\left[d_{\mathrm{avg}}(V,W)<\theta\right]\\ &= \sum_{W\in\mathcal{N}}\frac{1}{|\mathcal{U}|}\sum_{V\in\mathcal{U}}1\left\{d_{\mathrm{avg}}(V,W)<\theta\right\}\\ &\le \frac{|\mathcal{N}|\max_{W\in\mathcal{N}}N_{W,\theta}}{|\mathcal{U}|}.\end{aligned} \tag{C80}$$

In the second line, we define $\mathcal{N}$ be a minimal θ-covering net over $\mathcal{C}$ with respect to d_{avg}. Also, in the last line, we define $N_{W,\theta} \triangleq \sum_{V\in\mathcal{U}}1\{d_{\mathrm{avg}}(V,W)<\theta\}$ to be the number of $V\in\mathcal{U}$ that are θ-close to W in d_{avg}.

Now, we aim to upper bound $N_{W,\theta}$ by counting. We first note that $N_{W,\theta} \le \max_{V\in\mathcal{U}}N_{V,4\theta}+1$. This is because, by the definition of $N_{W,\theta}$, there exist $V_1,\ldots,V_{N_{W,\theta}}\in\mathcal{U}$ such that $d_{\mathrm{avg}}(V_i,W)<\theta, 1\le i\le N_{W,\theta}$. Then, for V_1 and any $V_i, 2\le i\le N_{W,\theta}$, we have

$$\begin{aligned}d_{\mathrm{avg}}(V_1,V_i) &\le d'_F(V_1,V_i) \le d'_F(V_1,W)+d'_F(V_i,W)\\ &\le 2d_{\mathrm{avg}}(V_1,W)+2d_{\mathrm{avg}}(V_i,W)<4\theta.\end{aligned} \tag{C81}$$

This means that there are at least $N_{W,\theta}-1$ elements of $\mathcal{U}$ that are (4θ)-close to V_1. Therefore, $N_{V_1,4\theta}\ge N_{W,\theta}-1$ and hence $N_{W,\theta}\le\max_{V\in\mathcal{U}}N_{V,4\theta}+1$.

Next, we upper bound $N_{V,4\theta}$ for any $V\in\mathcal{U}$. Recall that each $V\in\mathcal{U}$ is an oracle unitary of a Boolean function on $\{0,1\}^n$. We can represent it by $f_V(i)\in\{0,1\}, 1\le i\le 2^n$. Consider a different $V'\in\mathcal{U}$ corresponding to the Boolean function $f_{V'}$. If f_V and $f_{V'}$ differ on at least $\lceil 64\theta^2\cdot 2^n\rceil$ of the 2^n possible inputs $i\in[2^n]$, then the corresponding columns of the unitaries V and V' must also differ. In particular, in each of these columns, there will be a matrix element that is 1 for V but 0 for V'. This means that V and V' are 4θ apart from each other with respect to d_{avg}:

$$\begin{aligned}d_{\mathrm{avg}}(V,V') &\ge \frac{1}{2}\min_{e^{i\phi}\in U(1)}\|V-V'e^{i\phi}\|_F\\ &\ge \frac{1}{2\sqrt{d}}\min_{e^{i\phi}\in U(1)}\sqrt{64\theta^2\cdot 2^n|1-0\cdot e^{i\phi}|^2}=4\theta.\end{aligned} \tag{C82}$$

Therefore, all functions $f_{V'}$ corresponding to the $V'\in\mathcal{U}$ counted in $N_{V,4\theta}$ must differ from f_V on strictly less than $\lceil 64\theta^2\cdot 2^n\rceil$ of the 2^n inputs. This gives us

$$N_{V,4\theta} \le \sum_{k=0}^{\lceil 64\theta^2\cdot 2^n\rceil}\binom{2^n}{k}, \tag{C83}$$

where each term represents choosing k inputs where the output is different from f_V. The right-hand side can be further bounded as

$$\begin{aligned}\sum_{k=0}^{\lceil 64\theta^2\cdot 2^n\rceil}\binom{2^n}{k} &\le \left(\frac{e2^n}{\lceil 64\theta^2\cdot 2^n\rceil}\right)^{\lceil 64\theta^2\cdot 2^n\rceil}\\ &\le 2^{(64\theta^2\cdot 2^n+1)\log_2(e/64\theta^2)}.\end{aligned} \tag{C84}$$

Note that when $\theta = 1/16$, we have $64\theta^2 = 1/4$ and $64\theta^2 \log_2(e/64\theta^2) = \log_2(4e)/4 < 0.87$. Therefore, recalling that the set of all n-bit classical Boolean functions has size $|\mathcal{U}| = 2^{2^n}$, we obtain

$$\Pr\left[O_V > 1 - \frac{\theta^2}{4}\right] \leq |\mathcal{N}| 2^{-0.13\cdot 2^n + \log_2(4e)+1}, \tag{C85}$$

where the extra one in the exponent takes the one in $N_{W,\theta} \leq \max_{V\in\mathcal{U}} N_{V,4\theta} + 1$ into account.

Finally, we move on to bound $|\mathcal{N}|$. Similarly to Eq. (B102) in the state-learning case, since our learning algorithm is a polynomial-time algorithm that can only output circuit descriptions with size $\mathrm{poly}(n)$, we must have

$$|\mathcal{N}| \leq \mathcal{O}\left((1/\theta)^{\mathrm{poly}(n)}\right) = \mathcal{O}\left(2^{\mathrm{poly}(n)}\right). \tag{C86}$$

Thus we arrive at

$$\Pr\left[O_V > 1 - \frac{\theta^2}{4}\right] \leq \mathcal{O}\left(2^{\mathrm{poly}(n) - 0.13\cdot 2^n}\right) = \mathrm{negl}(n) \tag{C87}$$

and therefore

$$\begin{aligned}\Pr\left[\mathcal{D}^{|V\rangle}(\cdot) = 1\right] &\leq 1 - \frac{\theta^2}{4} + \mathrm{negl}(n) + (1-p) \\ &= \frac{8185}{8192} + \mathrm{negl}(n),\end{aligned} \tag{C88}$$

where we have used $\theta = 1/16$ and $p = 1 - 1/8192$.

Combining Eqs. (C77) and (C87), we have

$$\begin{aligned}&\left|\Pr_{U\in\mathcal{RF}}[\mathcal{D}^{|U\rangle}(\cdot) = 1] - \Pr_{V\in\mathcal{U}}[\mathcal{D}^{|V\rangle}(\cdot) = 1]\right| \\ &\geq \frac{4}{8192} - \mathrm{negl}(n) \geq \frac{1}{4096},\end{aligned} \tag{C89}$$

for large n. This contradicts the defining property of pseudorandom functions $\mathcal{RF}$ [Theorem 11, item (2)] under the assumption that RingLWE is hard. ■

Next, we invoke the stronger assumption that RingLWE cannot be solved by any subexponential-time quantum algorithms and show that learning unitaries composed of $G = \mathcal{O}(\log n \cdot \mathrm{polyloglog} n)$ gates is computationally hard.

Theorem 19 (Unitary-learning computational-complexity lower bound assuming subexponential hardness of RingLWE; restatement of lower bound in Theorem 6). Let $\lambda = l = \Theta(G)$, with $l \leq n$ being the security parameter. Let V be an l-qubit unitary consisting of $\mathcal{O}(l\mathrm{polylog}(l)) = \mathcal{O}(G\mathrm{polylog}(G))$ gates (or a depth $d = \mathcal{O}(\mathrm{polylog}(G))$ circuit) that implements a pseudorandom function in $\mathcal{RF}$. Such a unitary V exists by Corollary 3. Let $U = V \otimes I$, where the identity I is over the last $(n-l)$ qubits. Any quantum algorithm for learning a circuit description of the n-qubit unitary U to within $\epsilon \leq 1/64$ average-case distance d_{avg} with probability at least $2/3$ from $N = \mathrm{poly}(\lambda)$ queries to U must use $\exp(\Omega(\min\{G, n\}))$ time, if quantum computers cannot solve RingLWE in subexponential time.

Proof. With polynomial hardness of RingLWE replaced by subexponential hardness, Theorem 18 asserts that there are no subexponential (in l) quantum algorithms that can learn the l-qubit unitary V to within average-case distance $\epsilon < 1/64$ with success probability at least $2/3$. That is, any such learning algorithms must use time at least $\exp(\Omega(l)) = \exp(\Omega(\min\{G, n\}))$, since $l \leq n$. Meanwhile, a polynomial learning algorithm for the n-qubit unitary $U = V \otimes I$ can be used to learn the l-qubit unitary V in the same run time by discarding the last $(n-l)$ qubits, because the trace distance does not increase under such an operation and thus neither does d_{avg}. This implies the $\exp(\Omega(\min\{G, n\}))$ time lower bound for the n-qubit learning algorithm. ■

ALGORITHM 6. Distinguisher $\mathcal{D}$ for PRF.

Input: Query access to the unknown n-qubit G-gate unitary U.
An error parameter $\epsilon \in (0, 1/\sqrt{d})$.
Output: A unitary $\hat{U}$.
Let $t \leftarrow \left\lceil \log_2(1/(\epsilon\sqrt{d})) \right\rceil$.
Let $V_0 \leftarrow I$.
Let $\mathcal{N} \leftarrow$ an $(\epsilon/10^5)$-covering net of G-gate unitaries with respect to $\mathrm{d}_{\mathrm{avg}}$.
for $j \leftarrow 0$ *to* t **do**
  Let $p_j \leftarrow 2^j$.
  Let $\eta_j \leftarrow 8^{j-t-1}\delta$.
  Use the algorithm $\mathcal{A}$ in Appendix C 2 b with success probability $1-\eta_j$ and accuracy $1/(25000\sqrt{d})$ to find a candidate R_j in $\{(U_i V_j^\dagger)^{p_j} \mid U_i \in \mathcal{N}\}$ that is closest to $(UV_j^\dagger)^{p_j}$ in $\mathrm{d}_{\mathrm{avg}}$.
  Let $V_{j+1} \leftarrow R_j^{1/p_j} V_j$.
return $\hat{U} \leftarrow V_{t+1}$

Finally, we briefly show that learning becomes efficient when $G = \mathcal{O}(\log n)$. The idea is that with $\mathcal{O}(\log n)$ gates, there can only be at most $\mathcal{O}(\log n)$ qubits affected. Thus we can focus on these qubits and learning the unitary amounts to manipulating at most $2^{\mathcal{O}(\log n)} = \text{poly}(n)$ size matrices, which is efficient. Specifically, we have the following statement.

Proposition 13 (Learning unitaries with logarithmic circuit complexity efficiently; restatement of upper bound in Theorem 6). Let $\epsilon > 0$. Suppose that we are given N queries to an n-qubit unitary U consisting of $G = \mathcal{O}(\log n)$ two-qubit gates. There exists a learning algorithm that outputs a $\hat{U}$ such that $d_{\text{avg}}(U, \hat{U}) \leq \epsilon$ with probability at least $2/3$ using $\text{poly}(n, 1/\epsilon)$ queries and time.

Proof. We prove this by a learning algorithm similar to Proposition 7 via junta learning based on Choi states (Appendix C 2 b) as follows.

First, we prepare the Choi state of U by applying it to a maximally entangled state over $2n$ qubits, execute Algorithm 4, and postselect on the trivial qubits being in the state $|I\rangle\rangle$, as in Appendix C 2 b. This step uses $\text{poly}(n, 1/\epsilon)$ queries and time and gives us the postselected Choi state that is nontrivial on only $4G = \mathcal{O}(\log n)$ qubits. Then, we use the Pauli-tomography method as in Proposition 7 to learn a trace-distance approximation to the $4G$-qubit Choi state $|\hat{V}\rangle\rangle$ using $\text{poly}(n, 1/\epsilon)$ queries and time. We can enforce this approximation to be a valid Choi state by projecting it to the subspace spanned by $(A \otimes I)|\Phi\rangle$ and normalizing the projected state, where A is an arbitrary matrix and $|\Phi\rangle$ is the maximally entangled state. This can be done via a projector that is a $2^{4G} = \text{poly}(n)$ dimensional matrix. Finally, we calculate the corresponding unitary $\hat{V}$ and set $\hat{U} = \hat{V} \otimes I$. Note that this step is efficient as it only involves manipulating matrices of size $2^{4G} = \text{poly}(n)$. Since the trace distance between Choi states is equivalent to the average-case distance between the corresponding unitaries, this gives us a $\text{poly}(n, 1/\epsilon)$ learning algorithm for average-case unitary learning. ■

APPENDIX D: LEARNING PHYSICAL FUNCTIONS

As stated in the main text, learning classical functions that map variables controlling the input states and evolution to some property of the outputs is an alternative way of learning about nature. Learning such functions has long been a central task of statistics and, more recently, classical and quantum machine learning. However, the physical mechanism that gives rise to these functions has largely been overlooked for the convenience of mathematical abstraction.

In fact, we can formulate the physical mechanism underlying a classical function as an experimental procedure involving a unitary with bounded circuit complexity. Specifically, we consider the following general experimental setting:

(1) Given a set of ν variables $x \in [0, 1]^\nu$, we prepare a pure state that can depend on x in a fixed way.
(2) We evolve the state using a unitary $U(x; \{U_i\}_{i=1}^G, a)$ that contains at most G two-qubit gates $\{U_i\}_{i=1}^G$, which can be tuned arbitrarily, and any number of fixed unitaries, which can depend on x, according to a circuit architecture a in an architecture class A.
(3) We measure the output state with a fixed observable O and read out the expectation value as the function output.

We can, without loss of generality, absorb the state preparation into the unitary. Then, the experiment gives rise to the function

$$f(\cdot; \{U_i\}, a) : [0, 1]^\nu \ni x \mapsto f(x; \{U_i\}, a) = \left\langle 0^n \middle| U(x; \{U_i\}, a)^\dagger O U(x; \{U_i\}, a) \middle| 0^n \right\rangle. \tag{D1}$$

We define

$$\mathcal{F}_{G,A}^\nu = \{f(\cdot, \{U_i\}, a) : a \in A, U_i \in U(2^2), i = 1, \ldots, G\} \subseteq \mathbb{R}^{[0,1]^\nu} \tag{D2}$$

to be the function class given by a class of architectures A for G-gate unitaries. We call such functions *physical functions* and $\mathcal{F}_{G,A}^\nu$ the class of ν-variable physical functions with G gates and architectures A.

This experiment can also be understood as a QML problem, where we want to collect training data $\{x, f(x)\}$ to learn to approximate certain functions in a function class using the ansatz described above. Then, the tunable gates $\{U_i\}$ can be understood as variational or trainable parameters of our quantum neural network. We note that the data-encoding unitaries may simply use x as the angles for rotation, or it can also be arbitrarily complex (e.g., complex enough to implement a quantum random access memory [186] that prepares the amplitude encoding of the data) as long as it is not trainable. This encompasses the case in which the input data are classical descriptions of the input pure state. Also, the order of the data-encoding unitaries and the trainable unitaries can be arbitrary, thus accommodating data-reuploading strategies [187,188].

We will show that to approximate a certain class of functions well, we need a minimal number of samples to learn and a minimal number of gates G (Theorem 7). In particular, we consider the class of 1-bounded and 1-Lipschitz functions on $[0, 1]^\nu$, which can (up to equivalence classes) be represented by the unit ball $B^{1,\infty}$ in the Sobolev space $W^{1,\infty}_{[0,1]^\nu}$. We establish the following theorem, where the learning criterion is the standard one for learning real functions [78, Definition 16.1].

Theorem 20 (Sample and gate-complexity lower bounds on functions given by G-gate unitaries to approximate bounded Lipschitz functions; restatement of Theorem 7). Let $\mathcal{F}^{\nu}_{G,A} \subseteq \mathbb{R}^{[0,1]^{\nu}}$ be the function class given by an architecture class A of G two-qubit unitaries. Let $\epsilon \in (0,1)$ and let $l(|h(x)-y|)$ be a loss function, where l is a strictly increasing function with derivative larger than some positive constant on $[1,\infty)$. Suppose, for any 1-bounded and 1-Lipschitz function $f \in B^{1,\infty}$, that there exists an $h \in \mathcal{F}^{\nu}_{G,A}$ such that $\|f-h\|_{\infty} < \epsilon$. Then, the smallest training data size N such that there exists a learning algorithm $H : ([0,1]^{\nu},[0,1])^{N} \to \mathcal{F}^{\nu}_{G,A}$ that satisfies

$$\mathbb{P}_{S\sim P^{N}}\left\{\mathbb{E}_{(X,Y)\sim P} l(|H[S](X)-Y|) - \inf_{f\in\mathcal{F}^{\nu}_{G,A}} \mathbb{E}_{(X,Y)\sim P} l(|f(X)-Y|) \le \epsilon\right\} \ge 0.99, \tag{D3}$$

for any probability distribution P over $[0,1]^{\nu}\times[0,1]$ must be at least

$$N \ge \Omega\left(\frac{1}{\epsilon^{\nu}}\right). \tag{D4}$$

Moreover, we need at least

$$G \ge \tilde{\Omega}\left(\frac{1}{\epsilon^{\nu/2}}\right) \tag{D5}$$

two-qubit unitaries if A contains variable circuit structures or $G \ge \tilde{\Omega}(1/\epsilon^{\nu})$ if the circuit structure is fixed. The $\tilde{\Omega}$ for variable circuit structures hides logarithmic factors in ϵ as well as in the number of qubits n, while the $\tilde{\Omega}$ for fixed structure only hides logarithmic factors in ϵ.

This means that to approximate 1-bounded and 1-Lipschitz functions in ν-variables well to $\mathcal{O}(1/n^{D})$ accuracy, we need at least $\tilde{\Omega}(n^{\nu D/2})$ two-qubit unitaries and $\Omega(n^{\nu D})$ samples on which to train. Furthermore, $1/\exp(n)$ accuracy can only be achieved with exponential-size quantum circuits and exponentially many samples. This result establishes a limitation on the maximal efficiency of using parametrized quantum circuits to approximate functions, complementary to existing works on universal approximation theorems for parametrized quantum circuits [80–83].

The exponential dependence of the training data size N on the number of variables ν suggests that if one has an extensively large input vector (the length of which scales with n), then the number of samples and gates needed to approximate such functions is exponentially large. Moreover, if the variables are encoded using amplitude encoding [e.g., via quantum random access memory (QRAM)], which accommodates exponentially many variables (approximately 2^{n}), then the gate and sample requirement would grow double exponentially in $1/\epsilon$. This phenomenon, termed the curse of dimensionality, has also been established in the theory of classical neural networks [79, Chapter 3]. We show that it still exists in quantum machine learning.

This curse can be circumvented by introducing more structure or constraints on the function class. For example, if we constrain to Fourier-integrable functions, a ν-independent number of $\mathcal{O}(1/\epsilon^{2})$ parameters suffices for both classical [79, Theorem 3.9] and quantum [80] machine learning. However, the curse of dimensionality shows that many-variable 1-bounded and 1-Lipschitz functions are not physical [50,158] because nature cannot efficiently implement them.

In order to prove Theorem 20, we proceed in three steps. First, we show that the complexity of the function class $\mathcal{F}^{\nu}_{G,A}$ is limited by the number of gates G. Then, we prove that to approximate certain functions (1-bounded and 1-Lipschitz functions) well enough, the complexity must not be too small. Finally, we show that to learn a function class from data, the number of samples that we need is lower bounded by the complexity of the function class.

1. Circuit complexity and function complexity

The complexity of the function class $\mathcal{F}^{\nu}_{G,A}$, measured by the pseudodimension or fat-shattering dimension [189, 190], is limited by the number of trainable gates G and the size of the architecture class A. This is because, from the linearity of quantum mechanics, the function $f(x;\{U_i\},a)$ is a polynomial in the matrix elements of the trainable unitaries $\{U_i\}$ and the degree of this polynomial is limited by G. Following the idea of Ref. [77], we formalize this idea into the following lemma.

Lemma 27 (Functions given by G-gate unitaries are bounded-degree polynomials). Let $\mathcal{F}^{\nu}_{G,A}$ be the function class given by an architecture class A of G two-qubit unitaries. Then, there exists a set of functions $P^{\nu}_{G,A}$ in $32G+\nu$ real variables with size $|P^{\nu}_{G,A}| = |A|$ such that the following two properties hold:

(1) $\forall f \in \mathcal{F}^{v}_{G,A}$, there exist a $p \in P^{v}_{G,A}$ and an assignment of the first $32G$ variables such that p under this assignment is the same as f in the last v variables.
(2) Each $p \in P^{v}_{G,A}$ depends polynomially on the first $32G$ variables with degree at most $2G$.

Proof. We begin by noting that for any fixed architecture $a \in A$, the function $f(x, \{U_i\}, a)$ is a function of $32G + v$ real variables, where the first $32G = 2 \cdot 2^2 \cdot 2^2 \cdot G$ variables are the real and imaginary parts of the matrix elements of $\{U_i \in U(2^2)\}$ and the last v variables are the input data $x \in [0,1]^v$.

Next, we aim to prove that f is a bounded-degree polynomial in the unitary matrix elements. We follow the idea of Ref. [77, Lemma 1] and analyze the function $f(x, \{U_i\}, a)$ gate by gate. We note the following fact from linear algebra: for any state $|\psi\rangle$ and matrix U, the product $U|\psi\rangle$ is a state the amplitudes of which are linear combinations of the amplitudes of $|\psi\rangle$ and of matrix elements of U. Therefore, by applying $\{U_i\}_{i=1}^{G}$ and other unitaries that do not depend on $\{U_i\}$ sequentially according to the architecture a, we obtain a state the amplitude of which is a polynomial of the matrix elements of $\{U_i\}$ with degree at most G. Hence, the output scalar $\langle 0^n | U(x; \{U_i\}, a)^\dagger O U(x; \{U_i\}, a) | 0^n \rangle$ is a polynomial of the matrix elements of $\{U_i\}$ with degree at most $2G$. Fixing those $32G$ variables corresponds to fixing $\{U_i\}$ and thus specifying any particular function in $\mathcal{F}^{v}_{G,A}$ with this architecture a. Taking into account the dependence on x and gathering the function for each architecture $a \in A$, we arrive at the desired set of functions $P^{v}_{G,A}$ with $|P^{v}_{G,A}| = |A|$. ■

The fact that these functions are of bounded degree in the variables specifying the trainable unitaries implies an upper bound on the pseudodimension. We prove this with a reasoning analogous to Ref. [191] and Ref. [77, Theorem 2].

Proposition 14 (Pseudodimension upper bound for functions given by G-gate unitaries). Let $\mathcal{F}^{v}_{G,A}$ be the function class given by an architecture class A of G two-qubit unitaries. Then, the pseudodimension of $\mathcal{F}^{v}_{G,A}$ is at most $128G \log_2(16eG|A|)$.

Proof. Let $\{(x_i, y_i)\}_{i=1}^{m} \subseteq [0,1]^v \times \mathbb{R}$ be a set of data points satisfying that for any $C \subseteq \{1, \ldots, m\}$, there exists $f_C \in \mathcal{F}^{v}_{G,A}$ such that $f(x_i) - y_i \geq 0$ if and only if $i \in C$. That is, $\{(x_i, y_i)\}_{i=1}^{m}$ is pseudoshattered by $\mathcal{F}^{v}_{G,A}$. From Lemma 27, we know that there exists a set of functions P in $32G + v$ real variables with size $|P| = |A|$ such that for every C, there is a $p_C \in P$ and an assignment Ξ_C to the first $32G$ variable that satisfies $p_C(\Xi_C, x_i) - y_i \geq 0$ if and only if $i \in C$. This means that the set of functions $\{p(\cdot, x_i) - y_i : i = 1, \ldots, m, p \in P\}$ is a set of $m|A|$ polynomials of degree at most $2G$ in $32G$ real variables that has at least 2^m different consistent sign assignments [192]. Now, we invoke the following technical lemma.

Lemma 28 (Bounded-degree polynomials have a bounded number of consistent sign assignments [77,191, 193]). Let P be a set of real polynomials in v variables with $|P| \geq v$, each of degree at most $D \geq 1$. Then, the number of consistent sign assignments to P is at most $(8De|P|/v)^v$.

Thus we have

$$2^m \leq \left(\frac{8 \cdot 2G \cdot em|A|}{32G} \right)^{32G}. \tag{D6}$$

Taking the logarithm yields

$$m \leq 32G(\log_2(16eG|A|) + \log_2(m/(32G))). \tag{D7}$$

Let us first assume that $m \geq 32G$. If $\log_2(16eG|A|) \geq \log_2(m/(32G))$, then we have $m \leq 64G \log_2(16eG|A|)$. Otherwise, $\log_2(16eG|A|) < \log_2(m/(32G))$ and we have $m \leq 64G \log_2(m/(32G))$, which translates into $(\log_2(m/(32G))/m/(32G)) \geq \frac{1}{2}$. Thus $m/(32G) \leq 4$ and $m \leq 128G$. In both cases, we have $m \leq 128G \log_2(16eG|A|)$. If $m < 32G$, this is also true. Therefore, we have pseudodimension (by definition in Definition 6) at most $128G \log_2(16eG|A|)$. ■

A special case is for fixed circuit architecture $|A| = 1$, where we have pseudodimension at most $128G \log_2(16eG)$. On the other hand, if we allow variable structure of the trainable unitaries, then $|A| \leq \binom{n}{2}^G \leq n^{2G}$, and we have pseudodimension at most $128G \log_2(16eG) + 256G^2 \log_2(16eGn)$.

2. Function complexity and approximation power

Now that we know that the pseudodimension of such function class is upper bounded via the number of gates G, we can derive the minimal number of gates needed to obtain certain function-approximation power. Consider the class of 1-bounded and 1-Lipschitz functions on $[0,1]^v$, which can be represented by the unit ball $B^{1,\infty}$ in the Sobolev space $W^{1,\infty}_{[0,1]^v}$. In order to approximate these functions well, the pseudodimension (and also the fat-shattering dimension) of our function class cannot be too small.

Lemma 29 (Pseudo- or fat-shattering dimension and approximat ion power; variant of Ref. [190, Theorem 2.10] and Ref. [194, Theorem 4]). Let $\epsilon > 0$ and $\mathcal{F} \subseteq \mathbb{R}^{[0,1]^v}$ be a class of functions such that for any $f \in B^{1,\infty}$, there is an $h \in \mathcal{F}$ such that $\|f - h\|_\infty < \epsilon$. Then, the pseudodimension of $\mathcal{F}$ must be at least $1/(4\epsilon)^v$. The ϵ-fat-shattering dimension of $\mathcal{F}$ must be at least $1/(8\epsilon)^v$.

Proof. Let $m \in \mathbb{N}$, to be chosen later. Let $x_1, \ldots, x_M \in [0,1]^d$ be $M = (m+1)^\nu$ points on a cubic lattice such that $\|x_i - x_j\| \geq 1/m$ for all $i \neq j$. Let $y \in \mathbb{R}^M$ and we will now construct a smooth function that takes the y values at these lattice points. Specifically, we define

$$f(x) = \sum_{i=1}^{M} y_i \phi(m(x - x_i)), \tag{D8}$$

where $\phi(z) = \prod_{j=1}^{\nu} \varphi(z_j)$ and φ is a smoothed version of the triangular function that takes value 0 at $|z| \geq 1/2$ and value 1 at $z = 0$ and $|\partial_j \phi(z)| \leq C$ for any $C > 2$. In this way, we have $f(x_i) = y_i$ for all $1 \leq i \leq M$.

Next, for any $\alpha \in \{0,1\}^M$, set $y_i = \alpha_i/(Cm)$. This means that $|y_i| \leq 1/(Cm)$ and thus $f \in B^{1,\infty}$. Then, by assumption, there must be an $h \in \mathcal{F}$ such that $\|f - h\|_\infty < \epsilon$. In particular, we have $|f(x_i) - h(x_i)| = |y_i - h(x_i)| < \epsilon$ for all i.

Now, for the pseudodimension, we can choose m large enough (say, $m = \lfloor 1/(C^2\epsilon) \rfloor$) such that $\epsilon < 1/(2Cm)$. Then,

$$h(x_i) \geq \frac{1}{2Cm} \iff \alpha_i = 1, y_i = \frac{1}{Cm}. \tag{D9}$$

Therefore, by definition in Definition 6, $\{x_1, \ldots, x_M\}$ is pseudoshattered by $\mathcal{F}$ and thus the pseudodimension of $\mathcal{F}$ is at least $M = (m+1)^\nu \geq 1/(C^2\epsilon)^\nu$. Taking the limit $C \to 2$ yields the desired result.

For fat-shattering dimension, we can choose m large enough (say, $m = \lfloor 1/(C^3\epsilon) \rfloor$) such that $\epsilon < 1/(4Cm)$. Then,

$$\alpha_i = 1 \implies h(x_i) \geq \frac{1}{Cm} - \epsilon \geq \frac{1}{2Cm} + \epsilon \tag{D10}$$

and

$$\alpha_i = 0 \implies h(x_i) \leq \epsilon \leq \frac{1}{2Cm} - \epsilon. \tag{D11}$$

Therefore, by definition in Definition 7, $\{x_1, \ldots, x_M\}$ is ϵ-fat-shattered by $\mathcal{F}$, and thus the ϵ-fat-shattering dimension of $\mathcal{F}$ is at least $M = (m+1)^\nu \geq 1/(C^3\epsilon)^\nu$. Taking the limit $C \to 2$ yields the desired result. ■

3. Function complexity and sample complexity

Now, we aim to show that in order to learn a function class, the number of samples that we need is lower bounded by its complexity. In particular, we achieve this through the fat-shattering dimension.

Proposition 15 (Sample-complexity lower bound for real-valued functions by fat-shattering dimension; variant of Ref. [78, Theorem 19.5]). Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ with loss function $l_h(x,y) = l(|h(x) - y|)$. Suppose that l is an increasing (almost everywhere) differentiable function, i.e., $C = \inf_{t \geq 1} l'(t) > 0$. For $0 < \epsilon < 1, 0 < \delta \leq 0.01$, the smallest training data size N such that there exists a learning algorithm $H : (\mathcal{X}, [0,1])^N \to \mathcal{F}$ that satisfies

$$\mathbb{P}_{S \sim P^N} \left\{ \mathbb{E}_{(X,Y)\sim P} l(|H[S](X) - Y|) - \inf_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim P} l(|f(X) - Y|) \leq \epsilon \right\} \geq 1 - \delta \tag{D12}$$

for any probability distribution P over $\mathcal{X} \times [0,1]$ must be at least

$$N \geq C \frac{\mathrm{fat}(\mathcal{F}, \epsilon/\alpha) - 1}{32\alpha}, \quad \forall \alpha \in (0, 1/4). \tag{D13}$$

Note that this contains L_p loss functions as a special case, where $l_h(x,y) = |h(x) - y|^p$, and $l'(t) = pt^{p-1} \geq p = C$.

Proof. Similarly to the proof of Theorem 19.5 in Ref. [78], the idea is to reduce the problem to a discrete classification problem. Consider the class H_d of all functions mapping from a finite set $\{x_1, \ldots, x_d\} \subset \mathcal{X}$ to $\{0,1\}$. It is known that any learning algorithm for H_d has sample complexity at least $(d-1)/(32\epsilon)$ for small ϵ, δ ([78, Theorem 5.3]). Here, we show that, for any fixed α between 0 and 1/4, any learning algorithm for $\mathcal{F}$ to accuracy ϵ can be used to construct a learning algorithm for H_d to accuracy α/C, where $d = \mathrm{fat}(\mathcal{F}, \epsilon/\alpha)$. Then the proposition follows.

To see this, suppose that $\{x_1, \ldots, x_d\}$ is ϵ/α-shattered by $\mathcal{F}$, witnessed by $r_1, \ldots, r_d$. Suppose that L is a learning algorithm for $\mathcal{F}$. Then, we can construct a learning algorithm for H_d as follows. For each labeled example (x_i, y_i), assuming that y_i is deterministic given x_i, the algorithm passes to L the labeled example $(x_i, \tilde{y}_i)$, where $\tilde{y}_i = 2$ if $y_i = 1$ and $\tilde{y}_i = -1$ if $y_i = 0$. Let P be the original distribution on $\mathcal{X} \times \{0,1\}$ and let $\tilde{P}$ be the induced distribution on $\mathcal{X} \times \{-1, 2\}$. Then, suppose that L produces

a function $f : \mathcal{X} \to [0,1]$; the learning algorithm for H_d then outputs $h : \mathcal{X} \to \{0,1\}$, where $h(x_i) = 1$ if and only if $f(x_i) > r_i$. Thus we only need to prove that if $\mathbb{E}_{\tilde{P}} l_f - \inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l_g < \epsilon$, then $\mathbb{E}_P 1(h(x) \neq y) \leq \alpha/C$.

To show this, we claim that

$$\inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l_g = \inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l(|g(x) - \tilde{y}|) \leq \mathbb{E}_{\tilde{P}} \min\{l(|\hat{y} - \tilde{y}|), \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\}, \quad \text{(D14)}$$

where $r(x_i) = r_i$. This is because $\tilde{P}$ is concentrated on the shattered set. Then, for any assignment $\{\hat{y}_i \in \{r_i \pm \epsilon/\alpha\}, i = 1, \ldots, d\}$, there exists a $g \in \mathcal{F}$ such that $g(x_i) \geq \hat{y}_i$ if $\hat{y}_i = r_i + \epsilon/\alpha$ and $g(x_i) \leq \hat{y}_i$ if $\hat{y}_i = r_i - \epsilon/\alpha$. In particular, we consider the assignment of $\hat{y}_i$ such that $l(|\hat{y}_i - \tilde{y}_i|)$ is minimized. Then, there exists a function g^* satisfying the following property. If $\tilde{y}_i = -1$, then the minimizer is $\hat{y}_i = r_i - \epsilon/\alpha$ and we have $l(|g^*(x_i) - \tilde{y}_i|) \leq l(|\hat{y} - \tilde{y}_i|)$ since $\tilde{y}_i < g^*(x_i) \leq \hat{y}_i$. Similarly, if $\tilde{y}_i = 2$, then the minimizer is $\hat{y}_i = r_i + \epsilon/\alpha$ and we still have $l(|g^*(x_i) - \tilde{y}_i|) \leq l(|\hat{y} - \tilde{y}_i|)$ since $\hat{y}_i \leq g^*(x_i) \leq \tilde{y}_i$. Therefore, since $\tilde{y}_i$ and y_i is deterministic given x_i, we have found a single g^* such that $\mathbb{E}_{\tilde{P}} l(|g^*(x) - \tilde{y}|) \leq \mathbb{E}_{\tilde{P}} \min\{l(|\hat{y} - \tilde{y}|), \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\}$. Hence, the infimum over $g \in \mathcal{F}$ $\inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l(|g(x) - \tilde{y}|) \leq \mathbb{E}_{\tilde{P}} l(|g^*(x) - \tilde{y}|) \leq \mathbb{E}_{\tilde{P}} \min\{l(|\hat{y} - \tilde{y}|), \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\}$. Therefore,

$$\mathbb{E}_{\tilde{P}} l_f - \inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l_g \geq \mathbb{E}[l(|f(x) - \tilde{y}|) - \min\{l(|\hat{y} - \tilde{y}|), \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\}]. \quad \text{(D15)}$$

Consider the quantity inside the expectation, for $x = x_i$ with $y = 0$, $\tilde{y} = -1$, let $a = f(x_i) + 1, b = r_i - \epsilon/\alpha + 1$. Then, by Lagrange's mean-value theorem, there exists a c between a and b, such that this quantity can be written as

$$l(a) - l(b) = l'(c)(a - b) = l'(c)(f(x_i) - r_i + \epsilon/\alpha). \quad \text{(D16)}$$

If $l(|f(x) - \tilde{y}|) - \min\{l(|\hat{y} - \tilde{y}|), \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\} < C\epsilon/\alpha$, then

$$f(x_i) - r_i < \frac{\epsilon}{\alpha} \frac{C - l'(c)}{l'(c)} < 0 \quad \text{(D17)}$$

and we have $f(x_i) < r_i$ and $h(x_i) = 0 = y_i$. Similar arguments apply for $y = 1$. Thus,

$$\mathbb{E}_P 1(h(x) \neq y) \leq \tilde{P}[|f(x) - \tilde{y}|^p - \min\{|\hat{y} - \tilde{y}|^p, \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\} \geq C\epsilon/\alpha] \quad \text{(D18)}$$

$$\leq \frac{\alpha}{C\epsilon} \mathbb{E}_{\tilde{P}}[|f(x) - \tilde{y}|^p - \min\{|\hat{y} - \tilde{y}|^p, \hat{y} \in \{r(x) \pm \epsilon/\alpha\}\}] \quad \text{(D19)}$$

$$\leq \frac{\alpha}{C\epsilon} (\mathbb{E}_{\tilde{P}} l_f - \inf_{g\in\mathcal{F}} \mathbb{E}_{\tilde{P}} l_g) \leq \frac{\alpha}{C}. \quad \text{(D20)}$$

This completes the proof of Proposition 15. ■

With Proposition 14, Lemma 29, and Proposition 15, we can finally proceed to prove Theorem 20.

Proof of Theorem 20. To show the gate-number lower bound, note that from Proposition 14, $\text{Pdim}(\mathcal{F}^{\nu}_{G,A})$ is upper bounded by $128G\log_2(16eG) + 256G^2\log_2(16eGn)$ for variable circuit structures and by $128G\log_2(16eG)$ for fixed circuit structure. Meanwhile, from Lemma 29, we know that to approximate any ν-variable 1-bounded and 1-Lipschitz functions to ϵ error in $\|\cdot\|_\infty$, we must have $\text{Pdim}(\mathcal{F}^{\nu}_{G,A}) \geq 1/(4\epsilon)^{\nu}$ and $\text{fat}(\mathcal{F}^{\nu}_{G,A}, \epsilon) \geq 1/(8\epsilon)^{\nu}$. Therefore, for variable circuit structures, we have

$$1/(4\epsilon)^{\nu} \leq \text{Pdim}(\mathcal{F}^{\nu}_{G,A}) \leq 128G\log_2(16eG) + 256G^2\log_2(16eGn) \quad \text{(D21)}$$

and thus $G \geq \tilde{\Omega}(1/(\epsilon)^{\nu/2})$. Similarly, for fixed circuit structure, we have $G \geq \tilde{\Omega}(1/\epsilon^{\nu})$.

To show the sample-complexity lower bound, note that from Proposition 15, we have the sample complexity $N \geq C(\text{fat}(\mathcal{F}^{\nu}_{G,A}, \epsilon/\alpha) - 1)/32\alpha$. Setting $\alpha = 1/8$ and using the fat-shattering bound from Lemma 29, we arrive at $N \geq \Omega(1/\epsilon^{\nu})$. ■

APPENDIX E: DETAILS OF THE NUMERICAL EXPERIMENTS

In this appendix, we provide the implementation details of the numerical experiments presented in Sec. III. All simulations are conducted using JAX [195] and TensorCircuit [196], with the matrix product state back end.

We consider a large system size $n = 10\,000$ to illustrate the independence of the sample complexity from n. For

simplicity, we consider the setting in which the G gates used to generate the unknown target states are sampled from a discrete gate set of size 2. Here, we sample two Haar-random two-qubit gates to form this gate set. We also assume that the $n = 10\,000$ qubits form a one-dimensional line and that each gate only acts on neighboring qubits. The gate configurations are randomly sampled over the first four qubits [case (a)] or over all qubits [case (b)]. We further assume that the gate positions are known to the learning algorithm to accelerate the simulation. Of course, this assumption can be removed by enumerating over all $\binom{n'}{2}^G$ possible configurations (with $n' = 4$ or $\leq 2G$ the number of qubits acted upon nontrivially by the G gates) or by performing Algorithm 1. However, this will introduce an additional overhead to the implementation, which we want to avoid for the sake of these numerics.

For each G and sampled target state, we perform the learning algorithm detailed in Appendix B 1 and calculate the fidelity F between the output state and the target state. To accelerate the learning algorithm for different sample sizes, we first sample the Clifford classical shadows with the maximal sample size and then subsample the results for smaller sample sizes. In case (b), where the effective system size n' is larger (up to $2G = 12$), we adopt the shallow shadow modification of Clifford classical shadows [88,89]. Specifically, we replace the global Clifford rotation with a brickwork Clifford circuit of depth 10. We also replace Clifford gates with Haar random gates to reduce statistical fluctuations.

The procedure described above constitutes a single sweep over the entire G-N plane. We repeat this procedure independently 100 000 times for case (a) and 25 000 times for case (b). We record the resulting fidelity and calculate the average and median values for each G and N. The results are presented in Fig. 2.

We note that for each G, when the sample size is above the sample complexity, the reconstruction fidelity of most trials is exactly one up to machine precision. This is because we are using a discrete gate set, so the target states can be found unambiguously. Therefore, the sample-complexity lines for $F_{\text{med}} = 0.999, 0.9999, 0.99999$ coincide and we only plot the one for $F_{\text{med}} = 0.999$. Meanwhile, the line for averaged fidelity F changes with the value of F, because the average fidelity F takes into account those failing cases in which the fidelity can be close to zero.

APPENDIX F: PRELIMINARY RESULTS ON LEARNING BRICKWORK CIRCUITS

As stated in Sec. IV, an interesting circuit structure is the brickwork circuit, which is generated by repeatedly applying the following two layers of gates (suppose that n is even): (1) $U_{1,2} \otimes U_{3,4} \otimes \cdots \otimes U_{n-1,n}$ and (2) $U_{2,3} \otimes U_{4,5} \otimes \cdots \otimes U_{n-2,n-1}$, where $U_{i,j}$ denotes a two-qubit unitary acting on the ith and jth qubits. Here, we utilize the tools from unitary t-designs [130] to prove that the metric entropy of G-gate brickwork circuits is lower bounded by $\Omega(tn)$, if they can implement (approximate) unitary t-designs. Specifically, we have the following result.

Proposition 16 (Metric entropy lower bound of brickwork circuits). Let $U_G^{n,\text{brick}} \subseteq U(2^n)$ be the set of n-qubit unitaries that can be implemented with G-gate brickwork circuits. Suppose that the uniform distribution over $U_G^{n,\text{brick}}$ forms an ϵ-approximate t-design of $U(2^n)$ for some $\epsilon \in (0, 1/2)$. Then, we have

$$\log \mathcal{M}(U_G^{n,\text{brick}}, d_{\text{avg}}, \epsilon) \geq \Omega(tn). \tag{F1}$$

Proof. Suppose that $U_G^{n,\text{brick}}$ with the uniform distribution forms an ϵ-approximate t-design $\mathcal{E}$ of $U(2^n)$. We begin by recalling a moment bound for approximate unitary designs. ■

Lemma 30 (Moment bound of approximate unitary designs [33, proof of Lemma 1]). Suppose that $\mathcal{E}$ is an ϵ-approximate unitary t-design of $U(d)$. Then, for any unitary $V \in U(d)$, we have

$$\mathbb{E}_{U\sim\mathcal{E}}\left[|\text{tr}(U^\dagger V)|^{2t}\right] \leq (1+\epsilon)t!. \tag{F2}$$

Consequently, by Markov's inequality, we have the following lemma saying that a random element of a design is far apart from a fixed unitary with high probability.

Lemma 31 (Design elements are far away from any fixed unitary). Suppose that $\mathcal{E}$ is an ϵ-approximate unitary t-design of $U(d)$. Then, for any unitary $V \in U(d)$, we have

$$\begin{aligned}&\mathbb{P}_{U\sim\mathcal{E}}\left[\|U-V\|_F^2 \leq 2d(1-\Delta)\right]\\ &\quad\leq \mathbb{P}_{U\sim\mathcal{E}}\left[|\text{tr}(U^\dagger V)| \geq d\Delta\right] \leq \frac{1+\epsilon}{\Delta^{2t}}\frac{t!}{d^{2t}}.\end{aligned} \tag{F3}$$

Proof. To prove this, we use the above moment bound and Markov's inequality:

$$\begin{aligned}\mathbb{P}_{U\sim\mathcal{E}}\left[|\text{tr}(U^\dagger V)| \geq d\Delta\right] &= \mathbb{P}_{U\sim\mathcal{E}}\left[|\text{tr}(U^\dagger V)|^{2t} \geq d^{2t}\Delta^{2t}\right]\\ &\leq \frac{\mathbb{E}_{U\sim\mathcal{E}}\left[|\text{tr}(U^\dagger V)|^{2t}\right]}{d^{2t}\Delta^{2t}}\\ &\leq \frac{1+\epsilon}{\Delta^{2t}}\frac{t!}{d^{2t}}.\end{aligned} \tag{F4}$$

Furthermore, since $\|U-V\|_F^2 = 2d - 2\text{Re}[\text{tr}(U^\dagger V)] \leq 2d(1-\Delta)$ implies $|\text{tr}(U^\dagger V)| \geq \text{Re}[\text{tr}(U^\dagger V)] \geq d\Delta$, Lemma 31 follows. ■

Now, we apply a probabilistic argument by randomly choosing M IID unitaries $U_1, \ldots, U_M$ from $\mathcal{E}$. The probability that any two of them are far away from each other is given by

$$\mathbb{P}_{U_1,\ldots,U_M\sim\mathcal{E}}[\forall 1 \le i \ne j \le M, \|U_i - U_j\|_F^2 \ge 2d(1-\Delta)] \tag{F5}$$

$$= 1 - \mathbb{P}_{U_1,\ldots,U_M\sim\mathcal{E}}[\exists 1 \le i \ne j \le M, \|U_i - U_j\|_F^2 \le 2d(1-\Delta)] \tag{F6}$$

$$\ge 1 - \sum_{1\le i\ne j\le M} \mathbb{P}_{U_1,\ldots,U_M\sim\mathcal{E}}[\|U_i - U_j\|_F^2 \le 2d(1-\Delta)] \tag{F7}$$

$$\ge 1 - \frac{M(M-1)}{2}\frac{1+\epsilon}{\Delta^{2t}}\frac{t!}{d^{2t}} \tag{F8}$$

$$\ge 1 - M^2\frac{1+\epsilon}{\Delta^{2t}}\frac{t!}{d^{2t}}, \tag{F9}$$

where we have used the union bound in the first inequality and Lemma 31 in the last. Therefore, as long as we take $M < \sqrt{\lfloor(\Delta^{2t}/(1+\epsilon))(d^{2t}/t!\rfloor}$, we have

$$\mathbb{P}_{U_1,\ldots,U_M\sim\mathcal{E}}[\forall 1 \le i \ne j \le M, \|U_i - U_j\|_F^2 \ge 2d(1-\Delta)] > 0. \tag{F10}$$

Hence there must be at least one instance $V_1,\ldots,V_M \in \mathcal{E}$ such that $\|V_i - V_j\|_F^2 \ge 2d(1-\Delta)$ for any pair V_i, V_j. These unitaries form a $\sqrt{2d(1-\Delta)}$-packing net of $U_G^{n,\text{brick}}$ with respect to $\|\cdot\|_F$. Thus we have

$$\log \mathcal{M}(U_G^{n,\text{brick}}, \|\cdot\|_F, \sqrt{2d(1-\Delta)}) \ge \Omega\left(\frac{1}{2}\log\left\lfloor\frac{\Delta^{2t}}{1+\epsilon}\frac{d^{2t}}{t!}\right\rfloor\right). \tag{F11}$$

If we set $\sqrt{2d(1-\Delta)} = \sqrt{d}\epsilon$ (i.e., $\Delta = 1 - \epsilon^2/2$), we arrive at

$$\log \mathcal{M}(U_G^{n,\text{brick}}, d_F, \epsilon) \ge \Omega(tn). \tag{F12}$$

From the fact that quotienting out a global phase only changes the metric entropy by an additive $\Omega(\log(1/\epsilon))$ terms (Lemma 10) and the equivalence of d_F' and d_{avg} [Lemma 4, item (1)], we arrive at the desired result.

[1] K. Banaszek, M. Cramer, and D. Gross, Focus on quantum tomography, New J. Phys. **15**, 125020 (2013).

[2] R. Blume-Kohout, Optimal, reliable estimation of quantum states, New J. Phys. **12**, 043034 (2010).

[3] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Quantum state tomography via compressed sensing, Phys. Rev. Lett. **105**, 150401 (2010).

[4] Z. Hradil, Quantum-state estimation, Phys. Rev. A **55**, R1561 (1997).

[5] M. Mohseni, A. T. Rezakhani, and D. A. Lidar, Quantum-process tomography: Resource analysis of different strategies, Phys. Rev. A **77**, 032322 (2008).

[6] J. L. O'Brien, G. J. Pryde, A. Gilchrist, D. F. V. James, N. K. Langford, T. C. Ralph, and A. G. White, Quantum process tomography of a controlled-NOT gate, Phys. Rev. Lett. **93**, 080502 (2004).

[7] A. J. Scott, Optimizing quantum process tomography with unitary 2-designs, J. Phys. A: Math. Theor. **41**, 055308 (2008).

[8] I. L. Chuang and M. A. Nielsen, Prescription for experimental determination of the dynamics of a quantum black box, J. Mod. Opt. **44**, 2455 (1997).

[9] G. M. D'Ariano and P. Lo Presti, Quantum tomography for measuring experimentally the matrix elements of an arbitrary quantum operation, Phys. Rev. Lett. **86**, 4195 (2001).

[10] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum metrology, Phys. Rev. Lett. **96**, 010401 (2006).

[11] V. Giovannetti, S. Lloyd, and L. Maccone, Advances in quantum metrology, Nat. Photonics **5**, 222 (2011).

[12] C. Kokail, R. van Bijnen, A. Elben, B. Vermersch, and P. Zoller, Entanglement Hamiltonian tomography in quantum simulation, Nat. Phys. **17**, 936 (2021).

[13] J. Carrasco, A. Elben, C. Kokail, B. Kraus, and P. Zoller, Theoretical and experimental perspectives of quantum verification, PRX Quantum **2**, 010102 (2021).

[14] R. Levy, D. Luo, and B. K. Clark, Classical shadows for quantum process tomography on near-term quantum computers, Phys. Rev. Res. **6**, 013029 (2024).

[15] S. T. Merkel, J. M. Gambetta, J. A. Smolin, S. Poletto, A. D. Córcoles, B. R. Johnson, C. A. Ryan, and M. Steffen, Self-consistent quantum process tomography, Phys. Rev. A **87**, 062119 (2013).

[16] R. Blume-Kohout, J. King Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz, Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography, Nat. Commun. **8**, 14485 (2017).

[17] H.-Y. Huang, S. T. Flammia, and J. Preskill, Foundations for learning from noisy quantum experiments, arXiv:2204.13691.

[18] Z. Cai, R. Babbush, S. C. Benjamin, S. Endo, W. J. Huggins, Y. Li, J. R. McClean, and T. E. O'Brien, Quantum error mitigation, Rev. Mod. Phys. **95**, 045005 (2023).

[19] E. Van Den Berg, Z. K. Minev, A. Kandala, and K. Temme, Probabilistic error cancellation with sparse Pauli-Lindblad models on noisy quantum processors, Nat. Phys. **19**, 1116 (2023).

[20] R. S. Gupta, E. Van Den Berg, M. Takita, D. Ristè, K. Temme, and A. Kandala, Probabilistic error cancellation for dynamic quantum circuits, Phys. Rev. A **109**, 062617 (2024).

[21] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, Sample-optimal tomography of quantum states, IEEE Trans. Inf. Theory **63**, 5628 (2017).

[22] R. O'Donnell and J. Wright, in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2016), p. 899.

[23] J. Haah, R. Kothari, R. O'Donnell, and E. Tang, in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2023), pp. 363–390.

[24] J. Bardeen, L. N. Cooper, and J. Robert Schrieffer, Theory of superconductivity, Phys. Rev. **108**, 1175 (1957).

[25] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, Ann. Phys. **349**, 117 (2014).

[26] G. Torlai, C. J. Wood, A. Acharya, G. Carleo, J. Carrasquilla, and L. Aolitav, Quantum process tomography with unsupervised learning and tensor networks, Nat. Commun. **14**, 2858 (2023).

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

[28] G. Torlai, G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko, and G. Carleo, Neural-network quantum state tomography, Nat. Phys. **14**, 447 (2018).

[29] H. Zhao, G. Carleo, and F. Vicentini, Empirical sample complexity of neural network mixed state reconstruction, Quantum **8**, 1358 (2024).

[30] M. Favoni, A. Ipp, D. I. Müller, and D. Schuh, Lattice gauge equivariant convolutional neural networks, Phys. Rev. Lett. **128**, 032003 (2022).

[31] A. Apte, A. Ashmore, C. Cordova, and T.-C. Huang, Deep learning lattice gauge theories, arXiv:2405.14830.

[32] M. Medvidović and J. Robledo Moreno, Neural-network quantum states for many-body physics, arXiv:2402.11014.

[33] F. G. S. L. Brandão, W. Chemissany, N. Hunter-Jones, R. Kueng, and J. Preskill, Models of quantum complexity growth, PRX Quantum **2**, 030316 (2021).

[34] D. Poulin, A. Qarry, R. Somma, and F. Verstraete, Quantum simulation of time-dependent Hamiltonians and the convenient illusion of Hilbert space, Phys. Rev. Lett. **106**, 170501 (2011).

[35] I. M. Georgescu, S. Ashhab, and F. Nori, Quantum simulation, Rev. Mod. Phys. **86**, 153 (2014).

[36] B. Zeng, X. Chen, D.-L. Zhou, and X.-G. Wen, *Quantum Information Meets Quantum Matter* (Springer, New York, 2019).

[37] D. Malz, G. Styliaris, Z.-Y. Wei, and J. Ignacio Cirac, Preparation of matrix product states with log-depth quantum circuits, Phys. Rev. Lett. **132**, 040404 (2024).

[38] D. T. Stephen, A. Dua, A. Lavasani, and R. Nandkishore, Nonlocal finite-depth circuits for constructing symmetry-protected topological states and quantum cellular automata, PRX Quantum **5**, 010304 (2024).

[39] M. Foss-Feig, D. Hayes, J. M. Dreiling, C. Figgatt, J. P. Gaebler, S. A. Moses, J. M. Pino, and A. C. Potter, Holographic quantum algorithms for simulating correlated spin systems, Phys. Rev. Res. **3**, 033002 (2021).

[40] C. Schön, E. Solano, F. Verstraete, J. Ignacio Cirac, and M. M. Wolf, Sequential generation of entangled multiqubit states, Phys. Rev. Lett. **95**, 110503 (2005).

[41] Y. Huang and X. Chen, Quantum circuit complexity of one-dimensional topological phases, Phys. Rev. B **91**, 195143 (2015).

[42] S. Aaronson, in *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing* (Association for Computing Machinery, New York, NY, USA, 2018), pp. 325–338.

[43] V. Lyubashevsky, C. Peikert, and O. Regev, in *Advances in Cryptology—EUROCRYPT 2010: 29th Annual International Conference on the Theory and Applications of Cryptographic Techniques, French Riviera, May 30–June 3, 2010. Proceedings 29* (Springer, Berlin, Heidelberg, 2010), p. 1.

[44] O. Regev, On lattices, learning with errors, random linear codes, and cryptography, J. ACM **56**, 1 (2009).

[45] S. Arunachalam, A. Bredariol Grilo, and A. Sundaram, Quantum hardness of learning shallow classical circuits, SIAM J. Comput. **50**, 972 (2021).

[46] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren, Cryptographic hardness of learning halfspaces with Massart noise, Adv. Neural Inf. Process. Syst. **35**, 3624 (2022).

[47] D. Aggarwal, H. Bennett, Z. Brakerski, A. Golovnev, R. Kumar, Z. Li, S. Peters, N. Stephens-Davidowitz, and V. Vaikuntanathan, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2023), p. 1516.

[48] P. Ananth, A. Poremba, and V. Vaikuntanathan, in *Theory of Cryptography Conference* (Association for Computing Machinery, Orlando, FL, USA, 2023), p. 93.

[49] A. Anshu and S. Arunachalam, A survey on the complexity of learning quantum states, Nat. Rev. Phys. **6**, 59 (2024).

[50] K. Poland, K. Beer, and T. J. Osborne, No free lunch for quantum machine learning, arXiv:2003.14103.

[51] K. Sharma, M. Cerezo, Z. Holmes, L. Cincio, A. Sornborger, and P. J. Coles, Reformulation of the no-free-lunch theorem for entangled datasets, Phys. Rev. Lett. **128**, 070501 (2022).

[52] N. Yu and T.-C. Wei, Learning marginals suffices! arXiv:2303.08938.

[53] R. Vershynin, *High-Dimensional Probability*, Cambridge Series in Statistical and Probabilistic Mathematics Vol. 47 (Cambridge University Press, Cambridge, 2018).

[54] C. Bădescu and R. O'Donnell, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2021), p. 1398.

[55] T. Chen, S. Nadimpalli, and H. Yuen, in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2023), p. 1163.

[56] L. Yang and N. Engelhardt, The complexity of learning (pseudo)random dynamics of black holes and other chaotic systems, arXiv:2302.11013.

[57] A. Abbas, R. King, H.-Y. Huang, W. J. Huggins, R. Movassagh, D. Gilboa, and J. McClean, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2023), Vol. 36, pp. 44792–44819.

[58] M. Hinsche, M. Ioannou, A. Nietner, J. Haferkamp, Y. Quek, D. Hangleiter, J.-P. Seifert, J. Eisert, and R. Sweke, One *T* gate makes distribution learning hard, Phys. Rev. Lett. **130**, 240602 (2023).

[59] J. van Apeldoorn, Ph.D. thesis, University of Amsterdam, 2019.

[60] A. Ambainis, in *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2000), p. 636.

[61] A. Belovs, Variations on quantum adversary, arXiv:1504.06943.

[62] P. Hoyer, T. Lee, and R. Spalek, in *Proceedings of the Thirty-Ninth Annual ACM symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2007), p. 526.

[63] M.-D. Choi, Completely positive linear maps on complex matrices, Linear Algebra Appl. **10**, 285 (1975).

[64] A. Jamiołkowski, Linear transformations which preserve trace and positive semidefiniteness of operators, Rep. Math. Phys. **3**, 275 (1972).

[65] M. Jiang, S. Luo, and S. Fu, Channel-state duality, Phys. Rev. A **87**, 022310 (2013).

[66] D. W. Berry, A. M. Childs, and R. Kothari, in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (IEEE, 2015), p. 792.

[67] R. Cleve, D. Gottesman, M. Mosca, R. D. Somma, and D. Yonge-Mallo, in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2009), p. 409.

[68] J. Bavaresco, M. Murao, and M. Túlio Quintino, Unitary channel discrimination beyond group structures: Advantages of sequential and indefinite-causal-order strategies, J. Math. Phys. **63**, 042203 (2022).

[69] W.-T. Kuo, A. A. Akhtar, D. P. Arovas, and Y.-Z. You, Markovian entanglement dynamics under locally scrambled quantum evolution, Phys. Rev. B **101**, 224202 (2020).

[70] H.-Y. Hu, S. Choi, and Y.-Z. You, Classical shadow tomography with locally scrambled quantum dynamics, Phys. Rev. Res. **5**, 023027 (2023).

[71] M. C. Caro, H.-Y. Huang, N. Ezzell, J. Gibbs, A. T. Sornborger, L. Cincio, P. J. Coles, and Z. Holmes, Out-of-distribution generalization for learning quantum dynamics, Nat. Commun. **14**, 3751 (2023).

[72] X. Wang, Y. Du, Z. Tu, Y. Luo, X. Yuan, and D. Tao, Transition role of entangled data in quantum machine learning, Nat. Commun. **15**, 3716 (2024).

[73] Z. Yu, X. Zhao, B. Zhao, and X. Wang, Optimal quantum dataset for learning a unitary transformation, Phys. Rev. Appl. **19**, 034017 (2023).

[74] C.-Y. Lai and H.-C. Cheng, Learning quantum circuits of some *T* gates, IEEE Trans. Inf. Theory **68**, 3951 (2022).

[75] D. Pollard, *Convergence of Stochastic Processes*, Springer Series in Statistics (1984).

[76] M. J. Kearns and R. E. Schapire, Efficient distribution-free learning of probabilistic concepts, J. Comput. Syst. Sci. **48**, 464 (1994).

[77] M. C. Caro and I. Datta, Pseudo-dimension of quantum circuits, Quantum Mach. Intell. **2**, 14 (2020).

[78] M. Anthony, P. L. Bartlett, and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations* (Cambridge University Press, New York, 1999).

[79] P. Grohs and G. Kutyniok, *Mathematical Aspects of Deep Learning* (Cambridge University Press, New York, 2022).

[80] L. Gonon and A. Jacquier, Universal approximation theorem and error bounds for quantum neural networks and quantum reservoirs, arXiv:2307.12904.

[81] A. Pérez-Salinas, D. López-Núñez, A. García-Sáez, P. Forn-Díaz, and J. I. Latorre, One qubit as a universal approximant, Phys. Rev. A **104**, 012405 (2021).

[82] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Phys. Rev. A **103**, 032430 (2021).

[83] A. Manzano, D. Dechant, J. Tura, and V. Dunjko, Parametrized quantum circuits and their approximation capacities in the context of quantum machine learning, arXiv:2307.14792.

[84] J. M. Martyn, Z. M. Rossi, A. K. Tan, and I. L. Chuang, Grand unification of quantum algorithms, PRX Quantum **2**, 040203 (2021).

[85] A. Gilyén, Y. Su, G. Hao Low, and N. Wiebe, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2019), p. 193.

[86] Z. Yu, Q. Chen, Y. Jiao, Y. Li, X. Lu, X. Wang, and J. Z. Yang, Provable advantage of parameterized quantum circuit in function approximation, arXiv:2310.07528.

[87] H.-Y. Huang, Y. Liu, M. Broughton, I. Kim, A. Anshu, Z. Landau, and J. R. McClean, in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2024), p. 1343.

[88] C. Bertoni, J. Haferkamp, M. Hinsche, M. Ioannou, J. Eisert, and H. Pashayan, Shallow shadows: Expectation estimation using low-depth random Clifford circuits, Phys. Rev. Lett. **133**, 020602 (2024).

[89] T. Schuster, J. Haferkamp, and H.-Y. Huang, Random unitaries in extremely low depth, arXiv:2407.07754.

[90] S. Aaronson and D. Gottesman, Improved simulation of stabilizer circuits, Phys. Rev. A **70**, 052328 (2004).

[91] A. Montanaro, Learning stabilizer states by Bell sampling, arXiv:1707.04012.

[92] R. A. Low, Learning and testing algorithms for the Clifford group, Phys. Rev. A **80**, 052314 (2009).

[93] A. Rocchetto, Stabiliser states are efficiently PAC-learnable, Quantum Inf. Comput. **18**, 541 (2018).

[94] S. Grewal, V. Iyer, W. Kretschmer, and D. Liang, Efficient learning of quantum states prepared with few non-Clifford gates, arXiv:2305.13409.

[95] S. Grewal, V. Iyer, W. Kretschmer, and D. Liang, Efficient learning of quantum states prepared with few non-Clifford gates II: Single-copy measurements, arXiv:2308.07175.

[96] D. Hangleiter and M. J. Gullans, Bell sampling from quantum circuits, Phys. Rev. Lett. **133**, 020601 (2024).

[97] L. Leone, S. F. E. Oliviero, S. Lloyd, and A. Hamma, Learning efficient decoders for quasichaotic quantum scramblers, Phys. Rev. A **109**, 022429 (2024).

[98] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Efficient quantum state tomography, Nat. Commun. **1**, 149 (2010).

[99] O. Landon-Cardinal, Y.-K. Liu, and D. Poulin, Efficient direct tomography for matrix product states, arXiv:1002.4632.

[100] Y. Guo and S. Yang, Scalable quantum state tomography with locally purified density operators and local measurements, arXiv:2307.16381.

[101] E. Bernstein and U. Vazirani, in *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 1993), p. 11.

[102] M. Rötteler, in *International Symposium on Mathematical Foundations of Computer Science* (Springer, Berlin, Heidelberg, 2009), p. 663.

[103] S. Arunachalam, S. Bravyi, A. Dutt, and T. J. Yoder, in *18th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2023)*, edited by O. Fawzi and M. Walter, *Leibniz International Proceedings in Informatics (LIPIcs)*, (Schloss Dagstuhl—Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2023), Vol. 266, p. 3:1.

[104] G. Tóth, W. Wieczorek, D. Gross, R. Krischek, C. Schwemmer, and H. Weinfurter, Permutationally invariant quantum tomography, Phys. Rev. Lett. **105**, 250403 (2010).

[105] T. Moroder, P. Hyllus, G. Tóth, C. Schwemmer, A. Niggebaum, S. Gaile, O. Gühne, and H. Weinfurter, Permutationally invariant state reconstruction, New J. Phys. **14**, 105001 (2012).

[106] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter, Experimental comparison of efficient tomography schemes for a six-qubit state, Phys. Rev. Lett. **113**, 040503 (2014).

[107] C. Rouzé and D. Stilck França, Learning quantum many-body systems from a few copies, Quantum **8**, 1319 (2024).

[108] S. Aaronson, The learnability of quantum states, Proc. R. Soc. A: Math. Phys. Eng. Sci. **463**, 3089 (2007).

[109] H. Cai, Q. Ye, and D.-L. Deng, Sample complexity of learning parametric quantum circuits, Quantum Sci. Technol. **7**, 025014 (2022).

[110] H.-Y. Huang, R. Kueng, and J. Preskill, Predicting many properties of a quantum system from very few measurements, Nat. Phys. **16**, 1050 (2020).

[111] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, The randomized measurement toolbox, Nat. Rev. Phys. **5**, 9 (2023).

[112] J. Kunjummen, M. C. Tran, D. Carney, and J. M. Taylor, Shadow process tomography of quantum channels, Phys. Rev. A **107**, 042403 (2023).

[113] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, *et al.*, Quantum advantage in learning from experiments, Science **376**, 1182 (2022).

[114] H.-Y. Huang, S. Chen, and J. Preskill, Learning to predict arbitrary quantum processes, PRX Quantum **4**, 040337 (2023).

[115] M. C. Caro, Learning quantum processes and Hamiltonians via the Pauli transfer matrix, ACM Trans. Quantum Comput. **5**, 1 (2024).

[116] S. Chen, J. Li, B. Huang, and A. Liu, in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), p. 1205–1213.

[117] S. Chen, B. Huang, J. Li, A. Liu, and M. Sellke, in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE Computer Society, Santa Cruz, CA, USA, 2023), pp. 391–404.

[118] S. Chen, J. Cotler, H.-Y. Huang, and J. Li, in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), p. 574.

[119] K. Chen, Q. Wang, P. Long, and M. Ying, Unitarity estimation for quantum channels, IEEE Trans. Inf. Theory **69**, 5116 (2023).

[120] O. Fawzi, A. Oufkir, and D. Stilck França, Lower bounds on learning Pauli channels, arXiv:2301.09192.

[121] O. Fawzi, N. Flammarion, A. Garivier, and A. Oufkir, in *Proceedings of Thirty Sixth Conference on Learning Theory*, edited by G. Neu and L. Rosasco, *Proceedings of Machine Learning Research* (PMLR, July 12–15, 2023), Vol. 195, p. 1822.

[122] A. Oufkir, in *2023 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2023), pp. 1919–1924.

[123] J. van Apeldoorn, in *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*. (Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021).

[124] W. J. Huggins, K. Wan, J. McClean, T. E. O'Brien, N. Wiebe, and R. Babbush, Nearly optimal quantum algorithm for estimating multiple expectation values, Phys. Rev. Lett. **129**, 240501 (2022).

[125] J. van Apeldoorn, A. Cornelissen, A. Gilyén, and G. Nannicini, in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2023), p. 1265.

[126] H.-Y. Huang, Y. Tong, D. Fang, and Y. Su, Learning many-body Hamiltonians with Heisenberg-limited scaling, Phys. Rev. Lett. **130**, 200403 (2023).

[127] A. Dutkiewicz, T. E. O'Brien, and T. Schuster, The advantage of quantum control in many-body Hamiltonian learning, arXiv:2304.07172.

[128] H. Li, Y. Tong, T. Gefen, H. Ni, and L. Ying, Heisenberg-limited Hamiltonian learning for interacting bosons, npj Quantum Inf. **10**, 83 (2024).

[129] M. P. A. Fisher, V. Khemani, A. Nahum, and S. Vijay, Random quantum circuits, Annu. Rev. Condens. Matter Phys. **14**, 335 (2023).

[130] F. G. S. L. Brandao, A. W. Harrow, and M. Horodecki, Local random quantum circuits are approximate polynomial-designs, Commun. Math. Phys. **346**, 397 (2016).

[131] J. Haferkamp, Random quantum circuits are approximate unitary t-designs in depth $O\left(nt^{5+o(1)}\right)$, Quantum **6**, 795 (2022).

[132] A. R. Brown and L. Susskind, Second law of quantum complexity, Phys. Rev. D **97**, 086015 (2018).

[133] L. Susskind, Black holes and complexity classes, arXiv:1802.02175.

[134] L. Susskind, Computational complexity and black hole horizons, Fortschr. Phys. **64**, 24 (2016).

[135] A. Bouland, B. Fefferman, and U. Vazirani, in *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)* (Dagstuhl, Germany, 2020), Vol. 151, pp. 63:1–63:2.

[136] D. Stanford and L. Susskind, Complexity and shock wave geometries, Phys. Rev. D **90**, 126007 (2014).

[137] A. R. Brown, D. A. Roberts, L. Susskind, B. Swingle, and Y. Zhao, Complexity, action, and black holes, Phys. Rev. D **93**, 086006 (2016).

[138] J. Haferkamp, P. Faist, N. B. T. Kothakonda, J. Eisert, and N. Yunger Halpern, Linear growth of quantum circuit complexity, Nat. Phys. **18**, 528 (2022).

[139] Z. Li, Short proofs of linear growth of quantum circuit complexity, arXiv:2205.05668.

[140] M. Oszmaniec, M. Horodecki, and N. Hunter-Jones, Saturation and recurrence of quantum complexity in random quantum circuits, arXiv:2205.09734.

[141] J. Haferkamp, On the moments of random quantum circuits and robust quantum complexity, arXiv:2303.16944.

[142] C. W. Helstrom, Quantum detection and estimation theory, J. Stat. Phys. **1**, 231 (1969).

[143] V. V. Shende, S. S. Bullock, and I. L. Markov, in *Proceedings of the 2005 Asia and South Pacific Design Automation Conference* (Association for Computing Machinery, New York, NY, USA, 2005), p. 272.

[144] A. S. Holevo, Some estimates of the information transmitted by quantum communication channels, Probl. Inf. Transm. **9**, 177 (1973).

[145] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2016), p. 913.

[146] J. J. Vartiainen, M. Möttönen, and M. M. Salomaa, Efficient decomposition of quantum gates, Phys. Rev. Lett. **92**, 177902 (2004).

[147] Z. Ji, Y.-K. Liu, and F. Song, in *Advances in Cryptology—CRYPTO 2018: 38th Annual International Cryptology Conference, Santa Barbara, California, USA, August 19–23, 2018, Proceedings, Part III 38* (Springer-Verlag, Berlin, Heidelberg, 2018), p. 126.

[148] O. Goldreich, S. Goldwasser, and S. Micali, How to construct random functions, J. ACM **33**, 792 (1986).

[149] Z. Brakerski and O. Shmueli, in *Theory of Cryptography Conference* (Springer-Verlag, Berlin, Heidelberg, 2019), p. 229.

[150] A. Barenco, A. Berthiaume, D. Deutsch, A. Ekert, R. Jozsa, and C. Macchiavello, Stabilization of quantum computations by symmetrization, SIAM J. Comput. **26**, 1541 (1997).

[151] H. Buhrman, R. Cleve, J. Watrous, and R. de Wolf, Quantum fingerprinting, Phys. Rev. Lett. **87**, 167902 (2001).

[152] https://github.com/haimengzhao/bounded-gate-tomography

[153] Due to the concentration of Lipschitz functions on inputs drawn from the Haar measure, controlling this root-mean-square distance also leads to error bounds that hold with high probability over random input states.

[154] M. C. Caro, H.-Y. Huang, M. Cerezo, K. Sharma, A. Sornborger, L. Cincio, and P. J. Coles, Generalization in quantum machine learning from few training data, Nat. Commun. **13**, 4919 (2022).

[155] M. M. Wilde, *Quantum Information Theory* (Cambridge University Press, Cambridge, New York, 2013).

[156] J. Watrous, *The Theory of Quantum Information* (Cambridge University Press, 2018).

[157] A. A. Mele, Introduction to Haar measure tools in quantum information: A beginner's tutorial, Quantum **8**, 1340 (2024).

[158] T. Barthel and J. Lu, Fundamental limitations for measurements in quantum many-body systems, Phys. Rev. Lett. **121**, 080406 (2018).

[159] S. J. Szarek, in *Proceedings of Research Workshop on Banach Space Theory (Iowa City, Iowa, 1981)* (University of Iowa, Iowa City, Iowa, 1982), p. 169.

[160] C. Bŭdescu and R. O'Donnell, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (2021)*, pp. 1398–1411.

[161] V. N. Vapnik and A. Ya Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. **16**, 264 (1971).

[162] A. Banerjee, C. Peikert, and A. Rosen, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, Berlin, Heidelberg, 2012), p. 719.

[163] M. Naor and O. Reingold, Synthesizers and their application to the parallel construction of pseudo-random functions, J. Comput. Syst. Sci. **58**, 336 (1999).

[164] Y. Takahashi, S. Tani, and N. Kunihiro, Quantum addition circuits and unbounded fan-out, Quantum Inf. Comput. **10**, 872 (2010).

[165] N. Yu, R. Duan, and M. Ying, Five two-qubit gates are necessary for implementing the Toffoli gate, Phys. Rev. A **88**, 010304 (2013).

[166] P. W. Beame, S. A. Cook, and H. James Hoover, Log depth circuits for division and related problems, SIAM J. Comput. **15**, 994 (1986).

[167] G. Rosenthal, Query and depth upper bounds for quantum unitaries via Grover search, arXiv:2111.07992.

[168] P. Ananth, L. Qian, and H. Yuen, in *Annual International Cryptology Conference* (Springer-Verlag, Berlin, Heidelberg, 2022), p. 208–236.

[169] W. Kretschmer, L. Qian, M. Sinha, and A. Tal, in *Proceedings of the 55th Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York, NY, USA, 2023), p. 1589.

[170] T. Morimae and T. Yamakawa, in *Annual International Cryptology Conference* (Springer, 2022), p. 269.

[171] W. Kretschmer, in *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)* edited by M.-H. Hsieh,

Leibniz International Proceedings in Informatics (LIPIcs) (Schloss Dagstuhl—Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021), Vol. 197, p. 2:1.

[172] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie, in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, STOC '94* (Association for Computing Machinery, New York, USA, 1994), p. 273.

[173] G. Rosenthal, in *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (SIAM, 2024), p. 2508.

[174] H. Yuen, An improved sample complexity lower bound for (fidelity) quantum state tomography, Quantum **7**, 890 (2023).

[175] A. S. Holevo, Statistical decision theory for quantum systems, J. Multivar. Anal. **3**, 337 (1973).

[176] M. Boyer, G. Brassard, P. Høyer, and A. Tapp, Tight bounds on quantum searching, Fortschr. Phys.: Prog. Phys. **46**, 493 (1998).

[177] C. Gidney, Constructing large controlled NOTs, 2015, https://algassert.com/circuits/2015/06/05/Constructing-Large-Controlled-Nots.html. Accessed: 2023-08-20.

[178] M. M. Wolf, Quantum channels and operations—guided tour, 7 2012.

[179] J. Wright, Ph.D. thesis, Carnegie Mellon University, 2016.

[180] H.-Y. Huang, R. Kueng, and J. Preskill, Information-theoretic bounds on quantum advantage in machine learning, Phys. Rev. Lett. **126**, 190505 (2021).

[181] H. Flanders, On spaces of linear transformations with bounded rank, J. London Math. Soc. **s1–37**, 10 (1962).

[182] A. Eskenazis, P. Ivanisvili, and L. Streck, Low-degree learning and the metric entropy of polynomials, arXiv:2203.09659.

[183] G. M. Benedek and A. Itai, Learnability with respect to fixed distributions, Theor. Comput. Sci. **86**, 377 (1991).

[184] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Ltd, Hoboken, New Jersey, 1999).

[185] P. L. Bartlett, P. M. Long, and R. C. Williamson, in *Proceedings of the Seventh Annual Conference on Computational Learning Theory* (John Wiley & Sons, Ltd, Hoboken, New Jersey, 1994), p. 299.

[186] V. Giovannetti, S. Lloyd, and L. Maccone, Quantum random access memory, Phys. Rev. Lett. **100**, 160501 (2008).

[187] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, Data re-uploading for a universal quantum classifier, Quantum **4**, 226 (2020).

[188] M. C. Caro, E. Gil-Fuster, J. Jakob Meyer, J. Eisert, and R. Sweke, Encoding-dependent generalization bounds for parametrized quantum circuits, Quantum **5**, 582 (2021).

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

[190] M. M. Wolf, Mathematical foundations of supervised learning (2018), https://mediatum.ub.tum.de/doc/1723378/1723378.pdf.

[191] P. W. Goldberg and M. R. Jerrum, Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers, Mach. Learn. **18**, 131 (1995).

[192] A consistent sign assignment to a set of polynomials $p_1, \ldots, p_k$ is a vector $b \in \{-1, 0, 1\}^k$ such that there exists a set of input variables $z_1, \ldots, z_N \in \mathbb{R}$ such that $\mathrm{sgn}[p_i(z_1, \ldots, z_N)] = b_i$ for all $1 \leq i \leq k$..

[193] H. E. Warren, Lower bounds for approximation by nonlinear manifolds, Trans. Am. Math. Soc. **133**, 167 (1968).

[194] D. Yarotsky, Error bounds for approximations with deep ReLu networks, Neural. Netw. **94**, 103 (2017).

[195] J. Bradbury, R. Frostig, P. Hawkins, M. James Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: Composable transformations of PYTHON+NumPy programs, 2018.

[196] S.-X. Zhang, J. Allcock, Z.-Q. Wan, S. Liu, J. Sun, H. Yu, X.-H. Yang, J. Qiu, Z. Ye, Y.-Q. Chen, *et al.*, TensorCircuit: A quantum software framework for the NISQ era, Quantum **7**, 912 (2023).