Learning shallow quantum circuits

Hsin-Yuan Huang[†1,2,3], Yunchao Liu[†4], Michael Broughton[3], Isaac Kim[5], Anurag Anshu[6], Zeph Landau[4], and Jarrod R. McClean[3]

[1]California Institute of Technology
[2]Massachusetts Institute of Technology
[3]Google Quantum AI
[4]University of California, Berkeley
[5]University of California, Davis
[6]Harvard University

Abstract

Despite fundamental interests in learning quantum circuits, the existence of a computationally efficient algorithm for learning shallow quantum circuits remains an open question. Because shallow quantum circuits can generate distributions that are classically hard to sample from, existing learning algorithms do not apply. In this work, we present a polynomial-time classical algorithm for learning the description of any unknown n-qubit shallow quantum circuit U (with arbitrary unknown architecture) within a small diamond distance using single-qubit measurement data on the output states of U. We also provide a polynomial-time classical algorithm for learning the description of any unknown n-qubit state $|\psi\rangle = U|0^n\rangle$ prepared by a shallow quantum circuit U (on a 2D lattice) within a small trace distance using single-qubit measurements on copies of $|\psi\rangle$. Our approach uses a quantum circuit representation based on local inversions and a technique to combine these inversions. This circuit representation yields an optimization landscape that can be efficiently navigated and enables efficient learning of quantum circuits that are classically hard to simulate.

Contents

[†]Co-first author. Both authors contributed equally (listed in alphabetical order).

1 Introduction

The question of how to efficiently learn expressive classes of quantum states and circuits features prominently in quantum complexity theory, quantum algorithm design, and the experimental characterization of quantum devices. As a first step, one might consider the efficiency of learning shallow (constant depth) quantum circuits, where, to date, there has been no resolution despite considerable interest from a number of angles. From a complexity perspective, shallow quantum circuits are known to be more powerful than their classical counterparts [1–4], and under widely accepted complexity assumptions, sampling from the output distribution of shallow quantum circuits is classically hard to simulate [5–9]. This computational power provides the basis for quantum computational advantage with NISQ (noisy intermediate-scale quantum) devices and supports the quest for developing quantum algorithms based on learning parameterized shallow quantum circuits [10–24]. Within an experimental setting focused on coherent errors or gate calibration, characterizing a NISQ device can be modeled as learning what shallow quantum circuit the device is performing. Despite substantial interest in the question of learning shallow quantum circuits from these directions, to date, no polynomial time algorithm for learning shallow quantum circuits has been found. In this work, we introduce several efficient algorithms for two related tasks.

Theorem (Summary of main results)**.** *There are polynomial time algorithms for (1) learning the description of an unknown n-qubit shallow quantum circuit U (with arbitrary unknown architecture) within a small diamond distance, given access to U; (2) learning the description of an unknown n-qubit state $|\psi\rangle = U|0^n\rangle$ prepared by a shallow quantum circuit U (on a 2D lattice) within a small trace distance, given copies of $|\psi\rangle$.*

The main challenges in learning shallow quantum circuits are twofold. While foundational results in computational learning theory have established the efficient learnability of shallow classical circuits [25–27], these techniques may not apply to shallow quantum circuits, as these circuits can generate distributions with nontrivial correlations over the entire system that are classically hard to simulate [7–9]. Furthermore, even when the structure of a shallow quantum circuit is known up to parameterization, the optimization landscape for learning shallow quantum circuits is swamped with exponentially many suboptimal local minima [23]. The bad optimization landscape causes standard optimization methods, such as gradient descent algorithms and Newton methods, to fail in learning shallow quantum circuits.

To address these challenges, we consider a quantum circuit representation based on *local inversions*, which yields an optimization landscape that can be efficiently navigated. The local inversions disentangle qubits in each local region in a way that does not perturb the remaining system. We then show how these local inversions may be combined to build up the entire circuit without having to solve a computationally hard problem. Together, this new technique enables us to learn a natural class of quantum circuits that are classically hard to simulate.

1.1 Background

Learning shallow classical circuits Although the shallow quantum case has many conceptual challenges resulting from non-locality, the learnability of shallow classical circuits is a fundamental question in computational learning theory that has been well-studied and resolved in many cases. Learning constant-depth classical circuits with bounded fan-in gates (NC^0) is equivalent to learning juntas and can be performed in polynomial time from uniform samples [26]. In addition, quasi-polynomial time algorithms are known for learning constant-depth classical circuits with unbounded fan-in AND/OR gates (AC^0) [25], as well as mod p gates ($\mathsf{AC}^0[p]$) [27] in the PAC model. The

problem of learning shallow quantum circuits (QNC^0) and their output states are natural quantum analogs of learning Boolean circuits. As QNC^0 can be exponentially more powerful than AC^0 for some computational problems [4], it is natural to ask if shallow quantum circuits can be learned efficiently from random data samples.

Quantum machine learning When one parameterizes the gates in a quantum circuit, the parameterized quantum circuit forms an ML model, known as a *quantum neural network*, that can learn from data and make predictions on new inputs [10–16]. Since deep parameterized quantum circuits suffer from having *barren plateaus* in the optimization landscape [28, 29] and are challenging to implement on noisy quantum devices [30, 31], shallow quantum circuits have been subject to extensive study in recent years [17–24]. Various applications of learning shallow quantum circuits have been explored, ranging from compressing quantum circuits for implementing a unitary [16, 32–35], speeding up quantum dynamics [36–40], to learning generative models for sampling from predicted distributions [41–46]. While the optimization landscape for learning shallow quantum circuits is free from barren plateau [17], the landscape is swamped with exponentially many suboptimal local minima; see Section 10 and [23] for a study of this phenomenon. The presence of a large number of suboptimal local minima causes standard local optimization methods, such as gradient descent or Newton's method, to fail in learning parameterized shallow quantum circuits.

Efficient quantum tomography While quantum state and process tomography generally require exponential resources, performing tomography over some restricted families of states or processes can be made computationally efficient. Examples of such families include matrix product states [47–49], high-temperature Gibbs states [50–52], stabilizer states [53–56], quantum phase states [57], noninteracting Fermionic states [58], Clifford circuits with a small number of T gates [54, 56, 59], Pauli channels under structural assumptions [60–63], and interacting Hamiltonian dynamics [64–74] (see [75] for a recent survey). Most of these examples correspond to quantum circuit families that are classically easy to simulate [76–80]. In contrast, sampling from the output distribution of constant-depth quantum circuits is classically hard even when restricted to a 2D lattice [6, 81]. The experimental effort to characterize NISQ devices motivates the question of how to perform tomography for states and processes generated by shallow quantum circuits. While these states can be learned sample-efficiently using shadow tomography [82–84], no computationally efficient algorithms are known.

1.2 Our Results

We first focus on cases where one is given black-box access to the unknown unitary in (1) learning general shallow quantum circuits and (2) learning geometrically-local shallow quantum circuits. We then consider the more restricted model where one is only provided access to copies of an unknown state and focus on (3) learning quantum states prepared by geometrically-local shallow quantum circuits on 2-dimensional lattices.

1.2.1 Learning general shallow quantum circuits

Let U be an unknown n-qubit unitary generated by a shallow quantum circuit. The learning algorithm uses a randomized measurement dataset consisting of N samples about U [16, 39, 40, 85–88]. This dataset has been proposed as the classical shadow of U [85–87]. Each classical data sample specifies a random n-qubit product input state $|\psi_\ell\rangle = \bigotimes_{i=1}^n |\psi_{\ell,i}\rangle$ and a randomized Pauli measurement outcome $|\phi_\ell\rangle = \bigotimes_{i=1}^n |\phi_{\ell,i}\rangle$ on the output states $U|\psi_\ell\rangle$, where $|\psi_{\ell,i}\rangle, |\phi_{\ell,i}\rangle \in$

$\{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\}$ are single-qubit stabilizer states. Each data sample can be generated by a single query to U. Our goal is to learn U within a small diamond distance. The following results have the form of learning a circuit V acting on $2n$ qubits, such that $\|V - U \otimes U^\dagger\|_\diamond \leq \varepsilon$. Hence, V can be used to implement U by tracing out the n-qubit ancilla system.

Our first main result shows that one can learn U with a polynomial sample and computational complexity, with only the assumption that U is constant-depth (i.e., U has arbitrary unknown connectivity). Furthermore, the result applies even when the circuit generating U can have any number m of ancilla qubits used as working space and can have arbitrary two-qubit gates in SU(4) between any pair of the $n + m$ qubits so long as the resulting operation on the n system qubits is unitary. The learning algorithm is fully classical given the randomized measurement dataset.

Theorem 1 (Learning shallow quantum circuits; see Theorem 5)**.** *Given an unknown n-qubit unitary U generated by a constant-depth circuit over any two-qubit gates between any pair of qubits. One can learn a constant-depth circuit approximating U to diamond distance ε with high probability from $N = \mathcal{O}(n^2 \log(n)/\varepsilon^2)$ samples about U and $\mathrm{poly}(n)/\varepsilon^2$ classical running time.*

When the circuit is over a finite gate set, U can be learned to zero error with high probability from $N = \mathcal{O}(\log n)$ samples and $\mathrm{poly}(n)$ time.

1.2.2 Learning geometrically-local shallow quantum circuits

The algorithm for learning general shallow quantum circuits runs in polynomial time but with a large exponent. Furthermore, the depth of the learned circuit V, while constant, could be substantially greater than the depth of U. Motivated by the fact that most realistic quantum systems are geometrically local on a finite-dimensional lattice, it is natural to wonder if these aspects can be improved when learning geometrically-local quantum circuits on lattices. Next, we show that this is indeed the case.

See Theorem 6 for a related result on learning shallow circuits over any geometry represented by a bounded-degree graph.

Theorem 2 (Learning geometrically-local shallow circuits; see Theorem 7)**.** *Given an unknown n-qubit geometrically-local depth-d quantum circuit U over a k-dimensional lattice with $d, k = \mathcal{O}(1)$. One can learn a geometrically-local shallow circuit that approximates U to diamond distance ε with high probability from $N = \mathcal{O}(n^2 \log(n)/\varepsilon^2)$ classical data samples and either*

- *$\mathcal{O}(n^3 \log(n)/\varepsilon^2)$ classical running time with a learned circuit depth of $(k+1)4^{4(8kd)^k} + 1$.*
- *$(n/\varepsilon)^{\mathcal{O}((8kd)^{k+1})}$ classical running time with a learned circuit depth of $(k+1)(2d+1)+1$.*

When the circuit is over a finite gate set, U can be learned to zero error with high probability from $N = \mathcal{O}(\log n)$ samples and $\mathcal{O}(n \log(n))$ time with a learned circuit depth of $(k+1)(2d+1)+1$.

This shows that in the geometrically local setting, the learned circuit depth can achieve a linear blow-up. Furthermore, the learning algorithm works for $d = \mathrm{polylog}(n)$ depth circuits at the cost of quasipolynomial running time.

We remark that the more formal statement of the above theorem, which is labeled in this work as Theorem 7, can be straightforwardly generalized to a larger class of unitaries called *quantum cellular automata* (QCA), which play an important role in understanding quantum phases of matter [89–92]. These are unitaries that map any geometrically local operator to a geometrically local operator in the Heisenberg picture. For any such unitary, our proof technique applies without any modification, yielding an efficient algorithm for learning any QCAs. Interestingly, while shallow quantum circuits

are QCAs by definition, the converse statement is not necessarily true. For instance, shifting a set of qubits on a one-dimensional lattice trivially maps local operators to local operators. However, it is impossible to decompose this unitary into a geometrically local shallow quantum circuit [90]; see Ref. [91, 92] for other nontrivial examples of QCA. Therefore, our algorithm is applicable beyond shallow quantum circuits.

So far, we have been focusing on learning a shallow quantum circuit from a classical randomized measurement dataset. A natural question asks if further improvement is possible when we allow more general quantum query access to U. In the following, we show that by using quantum queries to U, an exponential improvement in query complexity is possible and this result is *asymptotically-optimal* in both time and query complexity for learning geometrically-local shallow circuits over finite gate sets. Surprisingly, quantum access also allows these circuits to be *with certainty*, dropping the familiar qualifier of high probability. The matching lower bounds stem from the need to query at least $\Omega(1)$ times to obtain any information about U and to write down the learned n-qubit circuit, which requires $\Omega(n)$ time.

Theorem 3 (Learning shallow circuits with quantum queries; see Theorem 8)**.** *An unknown n-qubit geometrically-local shallow quantum circuit U over a finite gate set can be learned to zero error with zero failure probability using $\Theta(1)$ queries to U and $\Theta(n)$ quantum computational time.*

1.2.3 Learning output states of geometrically-local shallow quantum circuits

Besides learning the n-qubit unitary U using input-output queries, it is natural to study the problem of learning a pure quantum state $|\psi\rangle$ prepared by a shallow quantum circuit U, i.e., $|\psi\rangle = U|0^n\rangle$. Here, instead of given access to U, we are only given copies of the pure state $|\psi\rangle$ as in quantum state tomography [47, 93]. As discussed in Section 1.1, most families of efficient learnable quantum states, such as matrix product states [47–49] and stabilizer states [53–56], correspond to quantum circuit families that are classically easy to simulate [77, 78]. In contrast, constant-depth quantum circuits are classically hard to simulate even when restricted to a 2D lattice [6, 7].

Learning $U|0^n\rangle$ from copies of $U|0^n\rangle$ has an incomparable difficulty to the earlier results because it has a less stringent requirement (learning an output state of U) but a more restricted access model (accessing copies of $U|0^n\rangle$ instead of U). While $|\psi\rangle = U|0^n\rangle$ can be learned from polynomially many copies [51, 94], the restricted access model makes the problem computationally more challenging, and the question of whether there exists a polynomial time algorithm remains open. We give an efficient algorithm when U is restricted to a 2D lattice.

Theorem 4 (Learning quantum states prepared by 2D shallow circuits; see Theorem 9)**.** *Given copies of an unknown pure state $|\psi\rangle$, with the promise that $|\psi\rangle = U|0^n\rangle$ for an unknown geometrically-local circuit U with depth d over a 2-dimensional lattice. One can learn a geometrically-local shallow circuit with depth $3d$ that prepares $|\psi\rangle$ to trace distance ε with high probability, using $2^{\mathcal{O}(d^2)}\cdot(n/\varepsilon)^{\mathcal{O}(1)}$ copies of $|\psi\rangle$, in time $\left(nd^3/\varepsilon\right)^{\mathcal{O}(d^3)}$. When the circuit U is over a finite gate set, $|\psi\rangle$ can be learned to zero error with high probability from $\mathcal{O}(\log(n))$ copies and $\mathcal{O}(n\log n)$ time.*

Similarly, this result applies to $d = \mathrm{polylog}(n)$ depth at the cost of quasipolynomial running time. The efficient learnability of quantum states prepared by a shallow quantum circuit acting on 3D lattices (or on more general geometries) remains a challenging and interesting open problem.

1.3 Discussion

Higher circuit depth In the general setting without geometric locality, we show that log-depth circuits require exponentially many quantum queries to learn within a small diamond distance (see

Prop. 3), which is proven by showing that log-depth circuits can implement Grover's oracle over 2^n elements and applying the Grover lower bound [95]. Therefore, our result for efficiently learning general constant-depth quantum circuits cannot be extended to much higher depth.

In the geometrically-local setting, Theorem 7 implies polynomial-time learnability for quantum circuits on a k-dimensional lattice up to $\log(n)^{1/k}$ depth, and quasi-polynomial time for up to $\mathrm{polylog}(n)$ depth. What structural assumptions allow us to efficiently learn quantum circuits beyond polylog-depth remains an important open question.

Worst-case vs average-case distance Motivated by the above discussion, it is natural to consider learning quantum circuits under weaker notions of distance, analogous to the classical notion of PAC learning. The standard notion of average-case distance in the literature [96, 97] is defined as the distance between output states when averaging over input states generated by Haar random unitaries. While learning polynomial-size quantum circuits to small average-case distance can be achieved with polynomial sample complexity [16, 39], the computational complexity of achieving a small average-case distance remains an open question.

In addition, Ref. [86] considered a weaker notion of an average-case error where the goal is to learn observables of the output state for random input states and showed that under this notion, any quantum circuit (even those with exponential depth) could be learned in quasi-polynomial time.

Verifying the learned shallow quantum circuit Our learning algorithm provably works under the promise that the unknown n-qubit channel $\mathcal{C}$ corresponds to a unitary $\mathcal{C}(\rho) = U\rho U^\dagger$ and the unitary U is generated by a shallow quantum circuit. This promise does not necessarily hold: U could be a deep quantum circuit that may or may not have a shallow quantum circuit implementation, and $\mathcal{C}$ may not be close to a unitary due to the noise in the quantum device. Even if there is no promise of $\mathcal{C}$, one can still bluntly apply our learning algorithm to learn an n-qubit channel $\mathcal{E}$ generated by a shallow quantum circuit. However, the learned circuit $\mathcal{E}$ is no longer guaranteed to be close to the true unknown channel $\mathcal{C}$. This raises the question of whether we can verify the learned circuit $\mathcal{E}$ or the promise on $\mathcal{C}$.

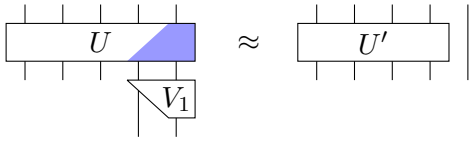
In Section 9, we give an efficient verification algorithm that outputs PASS if $\mathcal{E}$ is close to $\mathcal{C}$ in the average-case distance and $\mathcal{C}$ is close to unitary. The verification algorithm outputs FAIL if $\mathcal{E}$ is not close to $\mathcal{C}$. Because $\mathcal{E}$ is generated by a shallow quantum circuit, the verification algorithm only needs to use the classical dataset consisting of random input product states and randomized Pauli measurement outcomes on the outputs of $\mathcal{C}$.

Being able to verify the learned shallow quantum circuits is central to applications such as compressing quantum circuits for a known unitary. In this case, we have a known n-qubit unitary U that we know how to implement using a high-depth circuit. The goal is to learn a low-depth circuit that approximates U. If U does have a shallow circuit implementation, then our algorithm will learn a shallow circuit implementation for U. However, U may not have a shallow circuit implementation. In this case, the verification algorithm can tell us that our learning algorithm has failed. So far, we are using a simple verification algorithm based on a (weak) approximate local identity test, which only guarantees a small average-case distance. Whether more advanced verification schemes can be used to achieve stronger guarantees efficiently is an interesting question that requires further exploration.

2 Technical overview

Let U be an unknown n-qubit circuit of depth $d = \mathcal{O}(1)$. We consider the following two tasks: (1) Learn a constant-depth circuit $\hat{U}$ from random data samples from U or query access to U, such that U and $\hat{U}$ are close in diamond distance. (2) Learn a constant-depth circuit $\hat{U}$ from measuring copies of the n-qubit state $|\psi\rangle = U|0^n\rangle$, such that $\hat{U}|0^n\rangle$ and $U|0^n\rangle$ are close in trace distance.

A basic idea to learn U is to produce a guess $\hat{U}$ and check if $\hat{U}$ is close to U (i.e., $\hat{U}^\dagger \cdot U$ is close to identity). While the search space over $\hat{U}$ is exponentially large, the *locality* of shallow circuits allows us to search more efficiently. For example, in the following figure, we can find a small *local inversion* circuit V_1, that disentangles qubit 1 (the rightmost qubit), i.e., $UV_1 \approx U' \otimes I_1$. Here, the input wires are at the bottom, and the output wires are at the top; V_1 is applied before applying U.



(1)

This follows from a two-step argument. First, the existence of such a local inversion circuit is guaranteed by the locality of U, as undoing the gates in the backward lightcone (shaded blue region) of qubit 1 forms such a local inversion. Second, given a guess V_1, we develop an efficient procedure to check *approximate local identity*, i.e. $UV_1 \approx U' \otimes I_1$ for some $n-1$ qubit unitary U'. This allows us to find local inversions via brute force enumerate-and-test since the search space is small (as V_1 has depth d and is supported within a constant size region). Note that after this exhaustive process, we may find a list of valid local inversions. The "ground truth" local inversion compatible with the unique global inverse of the unitary is among them, but we do not know which one. Similarly, given copies of a state $|\psi\rangle = U|0^n\rangle$ we can find small local inversion circuits V_1 to disentangle qubit 1, $V_1|\psi\rangle \approx |\psi'\rangle \otimes |0\rangle_1$ for some $n-1$ qubit state $|\psi'\rangle$.

The above argument shows a procedure to efficiently learn local inversions for each qubit for both of our learning problems. The central question is whether this suffices to reconstruct the circuit and, if so, whether the reconstruction can be done efficiently. The main obstacle is that local inversions for each qubit are not unique, and two local inversions on neighboring qubits may not be consistent in the overlapping regions. Finding a consistent set of local inversions may require solving a constraint satisfaction problem that is computationally hard. Next, we show how to overcome this obstacle for learning U and $|\psi\rangle = U|0^n\rangle$.

2.1 Learning U to a small diamond distance

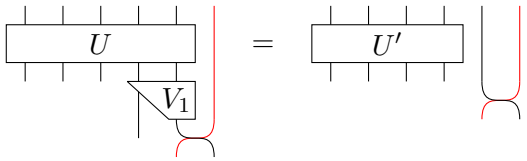
2.1.1 Sewing local inversions

Suppose we have learned a set of local inversions $\mathcal{C}_i$ for an unknown shallow quantum circuit U for each qubit i. Here, we show how to reconstruct the circuit using the learned local information. Surprisingly, the algorithm only requires an *arbitrary* element $V_i \in \mathcal{C}_i$ for each qubit i, without the need to search for the element compatible with the global inverse, which could require solving a complicated constraint satisfaction problem. The formal statements on this algorithmic technique are given in Section 5.2.1.

For simplicity, here we first assume all the local inversions are found exactly without any approximation. Take any $V_1 \in \mathcal{C}_1$, applying it to the unknown circuit U gives $UV_1 = U' \otimes I_1$, see Eq. (1), where we imagine qubit 1 to be the rightmost qubit and use a simple 1D geometry for illustration. This represents some progress: applying V_1 reduces the unknown n-qubit unitary U to an unknown

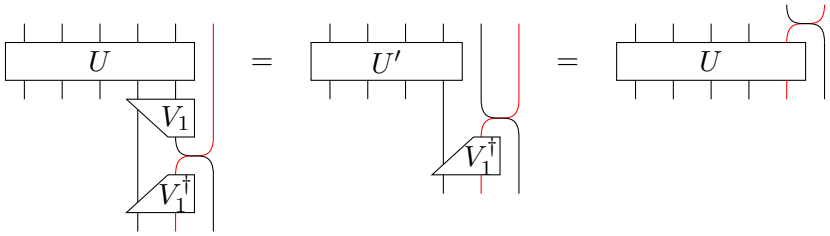
$(n-1)$-qubit unitary U' (note that U' may not be a shallow circuit). A natural thought is whether we can keep making this progress by applying local inversion on other qubits. The main issue here is that now the unitary has changed. For example, consider qubit 2 which is right next to qubit 1. Due to the fact that they have overlapping lightcones, some local inversion $V_2 \in \mathcal{C}_2$ may no longer work for the new circuit UV_1. Separately, we can attempt to find local inversion for qubit 2 with respect to this new circuit UV_1; however, doing so might disturb the progress we have made on qubit 1 and therefore requires coordinated effort across different qubits. This is exactly the type of constraint satisfaction problem that we want to avoid.

Here we introduce a general approach to keep making progress: the idea is to introduce a fresh ancilla qubit, swap it with qubit 1, and then *undo* the local inversion V_1. We show this in two steps: first, introduce a fresh ancilla qubit (red) and swap it with qubit 1,



(2)

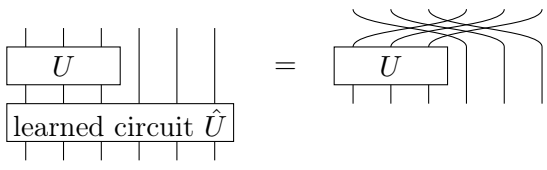
and then apply $V_1^\dagger$,



(3)

To explain the second equality of Eq. (3), note that without the swap operation, the above procedure is not doing anything (since we just perform some operation and undo it). In the second picture of Eq. (3), after experiencing $V_1^\dagger$, the red wire corresponds to the first output wire of U, but then it gets swapped out to the ancilla. Therefore, the overall effect is equivalent to performing a swap at the end after applying U.

The key reason that the above procedure is useful is because it *repairs* the circuit. This allows us to continue doing the same operation on qubit 2 because even though a lot of operations were applied before U (see the first picture in Eq. (3)), it is equivalent to as if nothing were applied before U (see the last picture in Eq. (3)); therefore we can similarly apply $V_2^\dagger$, swap with a new fresh qubit, and V_2 before U, achieving the effect of swapping qubit 2 at the end. Repeating the above procedure for all qubits, we have learned a circuit $\hat{U}$ acting on $2n$ qubits that satisfies



(4)

which implies that $\hat{U} = S \cdot (U \otimes U^\dagger)$, where S denotes the global swap operation between the system and ancilla qubits. To implement U using the learned circuit, on input ρ we initialize an ancilla register with some arbitrary state (say $|0^n\rangle$), apply $S \cdot \hat{U}$ and trace out the ancilla register, and the output state equals $U\rho U^\dagger$. We can use a similar procedure to implement $U^\dagger$. Thus, the above procedure simultaneously learns to implement U and $U^\dagger$, using access only to U.

Finally, we remark that the learned circuit $S \cdot \hat{U}$ is shallow. To see this, note that $S = \text{SWAP}^{\otimes n}$ is depth-1. $\hat{U}$ consists of unitaries of the form $W_i := V_i \cdot \text{SWAP} \cdot V_i^\dagger$ that are *local*: each of them supports on the lightcone of qubit i, as well as an extra ancilla qubit. Therefore we can implement non-overlapping W_is simultaneously, and all of the W_is can be stacked into a constant number of layers since, at most, a constant number of qubits share overlapping lightcones.

To achieve the optimal query and time complexity of $\Theta(1), \Theta(n)$ for learning geometrically-local shallow quantum circuits over finite gate sets in Theorem 3, we present a quantum learning algorithm that finds the exact local inversions for all n qubits with zero failure probability by querying U for only $\mathcal{O}(1)$ times. This surprising scaling is achieved by combining a few ideas: (a) coloring the geometry described by a bounded-degree graph, (b) decoupling the n-qubit unitary U into $\mathcal{O}(n)$ few-qubit channels based on the coloring, and (c) designing a tournament to perfectly distinguish between two classes of few-qubit quantum channels: those that form an exact local identity versus those that do not. The tournament uses the perfect distinguishability of certain pairs of CPTP maps shown in [98], where we design the few-qubit channels to ensure perfect distinguishability. Then, the learning algorithm finds a good order to sew the local inversions to produce a constant-depth circuit implementation for the unknown constant-depth n-qubit circuit U.

2.1.2 Sewing Heisenberg-evolved Pauli operators

Next, we describe a simpler technique based on directly sewing the Heisenberg-evolved Pauli operators $U^\dagger P_i U$ (P_i is a single-qubit Pauli acting on qubit i) and discuss how it is closely related to local inversion. Section 5.2.2 provides a detailed discussion of this technique.

We first describe how to learn the Heisenberg-evolved Pauli operators. Because U is a shallow quantum circuit, each operator $U^\dagger P_i U$ acts on a constant number of qubits. The few-qubit observable $U^\dagger P_i U$ can be reconstructed from the randomized measurement dataset. Let the random input product state be $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$, where $|\psi_i\rangle$ is a random one-qubit stabilizer state. Because each qubit in the output state is measured in a random X, Y, Z basis with equal probability, we will measure P_i on the output state $U|\psi\rangle\langle\psi|U^\dagger$ with probability 1/3. This allows us to estimate $\langle\psi|U^\dagger P_i U|\psi\rangle$. Then, we show that we can efficiently reconstruct $U^\dagger P_i U$ from a small number of different random input states.

After learning the $3n$ Heisenberg-evolved Pauli operators $U^\dagger P_i U$, we present a direct approach for sewing them into a circuit. This approach uses the identity $\text{SWAP} = \frac{1}{2}\sum_{P\in\{I,X,Y,Z\}} P \otimes P$. Let S_i be the SWAP gate acting on the i-th system qubit and the i-th ancilla qubit, let $S = \otimes_{i=1}^n S_i$ be the global swap between system and ancilla, and let $W_i := U^\dagger S_i U = \frac{1}{2}\sum_{P\in\{I,X,Y,Z\}} U^\dagger P_i U \otimes P, \forall i = 1, \ldots, n$. From the previous technique for sewing local inversion, we have proven the identity

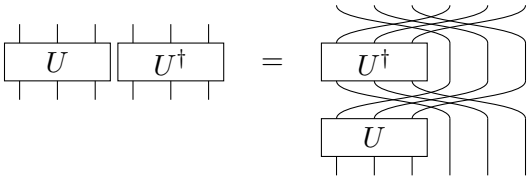
$$U \otimes U^\dagger = S \cdot \prod_{i=1}^{n} \left(V_i \cdot S_i \cdot V_i^\dagger \right), \tag{5}$$

where V_i satisfies $UV_i = U'^{(i)} \otimes I_i$ is an arbitrary exact local inversion on qubit i. We can see that

$$V_i \cdot S_i \cdot V_i^\dagger = U^\dagger U V_i \cdot S_i \cdot V_i^\dagger U^\dagger U = U^\dagger S_i U = W_i \Longrightarrow U \otimes U^\dagger = S \cdot \prod_{i=1}^{n} W_i = S \cdot \prod_{i=1}^{n} \left(U^\dagger S_i U \right). \tag{6}$$

The new equation can also be seen by itself: simply cancel U with $U^\dagger$ in the product so that the

right-hand side becomes $SU^\dagger SU$, and observe that



(7)

As we can see, the Heisenberg-evolved Pauli operators can be directly sewn into $U \otimes U^\dagger$.

This outlines the following procedure to learn U: first learn the Heisenberg-evolved Pauli operators $\{U^\dagger P_i U\}_{i=1}^n$, combine them to form $\{W_i\}_{i=1}^n$ according to $W_i = \frac{1}{2}\sum_{P\in\{I,X,Y,Z\}} U^\dagger P_i U \otimes P_i$, and reconstruct the circuit using $\{W_i\}_{i=1}^n$. Note that each W_i acts on a constant number k of qubits and can be directly compiled into a circuit of depth $2^{O(k)}$. To further optimize the depth of the learned circuit, notice that each W_i has the form $W_i = U^\dagger S_i U = V_i S_i V_i^\dagger$, i.e., it can be represented by a depth-$(2d+1)$ circuit. We can find such a representation for W_i by brute-force enumerating all depth-$(2d+1)$ circuits acting on k qubits, and the learned circuit has the same form as in Section 2.1.1. This thus provides a simpler framework for learning an unknown shallow quantum circuit U using a classical dataset containing random samples about U.

To prove Theorem 1 and 2 on learning general and geometrically-local shallow quantum circuits, we combine this framework with some additional ideas on (a) coloring the k-dimensional lattices to ensure all qubits with the same color has nonoverlapping lightcone, (b) truncating small Fourier coefficients to ensure the learned observables acts only on qubits in the support of the true observables, (c) compiling the Heisenberg-evolved Pauli operator when over a finite gate set, and (d) finding a good order to sew the Heisenberg-evolved Pauli operators into a short-depth circuit.

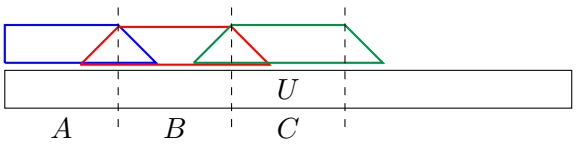
2.2 Learning $U|0^n\rangle$ to a small trace distance

Next, we discuss how to learn a quantum state $|\psi\rangle = U|0^n\rangle$ prepared by a shallow circuit U, given copies of $|\psi\rangle$. While this problem appears to be simpler (we need to learn $U|0^n\rangle$ instead of the entire U), the weaker access model (we only have access to the output of U for the all-zero input state $|0^n\rangle$) poses new fundamental challenges. In particular, we can learn local inversions V_i that give $V_i U|0^n\rangle = |\psi'\rangle \otimes |0\rangle_i$ instead of the much stronger $UV_i = U' \otimes I_i$, and the previous approach of "keep making progress by swapping ancilla qubits" does not seem to work.

Here, we address these challenges by developing new techniques tailored to a 2D lattice. The main idea is to *disentangle* the state into many 1D-like states that are easy to learn by leveraging the fact that 1D constraint satisfaction problems can be efficiently solved.

2.2.1 Disentangling a 2D quantum state

Our starting point is the simpler problem of learning a state $|\psi\rangle = U|0^n\rangle$, with the promise that U is a shallow circuit (white box) acting on a 1D lattice:

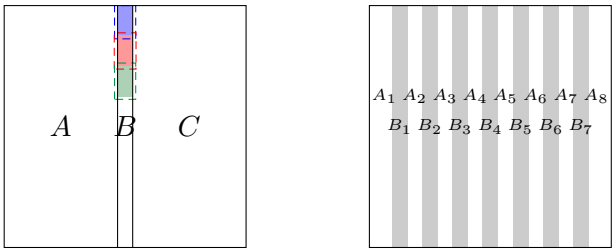


(8)

Let A, B, and C be contiguous regions of constant size. We can find a set of local inversions $\mathcal{C}_A$ for A by enumerating over circuits acting on the lightcone of A (blue shape). The question is how to combine different local inversions into a circuit. The key observation is that two neighboring local

inversions can be merged together if they are "consistent", i.e., sharing the same gates where they overlap. For example, some $V_A \in \mathcal{C}_A$ (blue) and $V_B \in \mathcal{C}_B$ (red) can be merged into a larger circuit of the same depth V_{AB} if they share the same gates in the overlapping region (intersecting triangle); the merged circuit V_{AB} satisfies $V_{AB}|\psi\rangle = |\psi'\rangle \otimes |0\rangle_{AB}$. This defines a constraint satisfaction problem: we need to find a local inversion for each region such that neighboring local inversions are consistent. Such a solution must exist (since the "ground truth" local inversions satisfy these constraints), and we can efficiently find such a solution by simple dynamic programming in time $\mathcal{O}(n|\mathcal{C}|^2)$ where $|\mathcal{C}|$ denotes the maximum number of local inversions for a small region. This gives a circuit V that satisfies $V|\psi\rangle = |0^n\rangle$, so the state $|\psi\rangle$ can be prepared by $|\psi\rangle = V^\dagger|0^n\rangle$.

From this perspective, generalizing this approach to 2D may be a difficult task since constraint satisfaction problems on 2D lattices are NP-hard in general. We address this challenge using an additional insight: instead of solving the constraint satisfaction problem directly in 2D, we first use the 1D argument to disentangle the 2D state.



(9)

The LHS of (9) shows a quantum state $|\psi\rangle$ prepared by a depth-d circuit acting on a 2D lattice, divided into three regions A, B, and C. A well-known fact about these states is that they have *finite correlation length*: if the width of B is sufficiently large (say $5d$), then the mutual information between A and C is zero, i.e. the reduced density matrix of $\rho = |\psi\rangle\langle\psi|$ on AC satisfies $\rho_{AC} = \rho_A \otimes \rho_C$. This fact itself does not simplify the problem because A and C are both entangled with B. However, if for some reason we have $\rho_B = |0\rangle\langle 0|_B$, then this would force ρ_A and ρ_C to be pure states and not entangled with any outside qubits.

But this is exactly what we can achieve using the 1D argument: we can learn local inversions for a small piece of B (shaded blue) by finding circuits acting on a slightly larger region (dotted blue). We can do this for contiguous small regions (here, the blue, red, and green regions play exactly the same role as in (8)), and by repeating the 1D argument we can find a depth-d circuit V acting on a region slightly larger than B, such that $\mathrm{Tr}_{AC}(V|\psi\rangle\langle\psi|V^\dagger) = |0\rangle\langle 0|_B$. After applying V, the state becomes $|\phi\rangle_A \otimes |0\rangle_B \otimes |\phi\rangle_C$ for some unknown pure states $|\phi\rangle_A$, $|\phi\rangle_C$.

Finally, note that this argument can be repeated horizontally across the entire system; overall, we can learn a depth-d circuit V such that $V|\psi\rangle$ has the form of RHS in (9). Here, all the shaded B regions are inverted and in the state $|0\rangle$. Each of the white regions is in a pure state and disentangled with each other. Now, the problem is reduced to learning each of the states $|\phi\rangle_{A_i}$ on the white regions separately. To prepare $|\psi\rangle$, we first prepare $(\otimes_i |\phi\rangle_{A_i}) \otimes |0\rangle_B$, then apply $V^\dagger$.

2.2.2 Learning finite correlated states in 1D

Here we address the final step of learning the 1D-like states $|\phi\rangle_{A_i}$. The main challenge here is that the previous argument in (8) is not immediately applicable: we do not have the guarantee that $|\phi\rangle_{A_i}$ is prepared by a shallow circuit acting on $|0\rangle_{A_i}$. Instead, what we know is that the global state $(\otimes_i |\phi\rangle_{A_i}) \otimes |0\rangle_B$ is prepared by a depth-$2d$ circuit acting on $|0\rangle_{AB}$, because it equals to $V|\psi\rangle$.

Our starting point is to observe the following structure of the state $|\phi\rangle_{A_i}$: it can be prepared by a depth-$2d$ circuit acting on A_i as well as some ancilla qubits A_i^L and A_i^R (see Fig. 5 for an

illustration). To see this, recall that $|\phi\rangle_{A_i}$ is part of a state that is prepared by a depth-$2d$ circuit. Now, imagine that we *undo* all the gates in that circuit, except for those in the *backward lightcone* of A_i. This procedure does not affect the state on A_i, and the resulting circuit (denoted as W_i) has exactly the same shape as in Fig. 5, where A_i^L, A_i^R both have width $2d$. We then develop an algorithm to learn such a depth-$2d$ circuit to prepare $|\phi\rangle_{A_i}$. This problem is different from (8) in nature due to the existence of ancilla qubits. However, its simple 1D structure allows us to develop a similar argument by solving a 1D constraint satisfaction problem. This implies that we can learn a depth-$2d$ circuit to prepare the entire system in RHS of (9). Thus the total learned circuit depth to prepare $|\psi\rangle$ equals $3d$ (see Claim 2 of Theorem 9).

In addition, we give a separate argument showing that each of the disentangled states $|\phi\rangle_{A_i}$ in RHS of (9) can be prepared with a 1D circuit of depth $2^{\mathcal{O}(d^2)}$ without any ancilla qubits. This implies an algorithm where the learned circuit for preparing $|\psi\rangle$ has depth $2^{\mathcal{O}(d^2)}$ and does not use ancilla qubits (see Claim 3 of Theorem 9).

Finally, note that throughout Section 2.2.1 and 2.2.2 we have been working with a simple setting with a finite gate set, which allows each step in the above argument to be performed *exactly* without any approximation error. Generalizing these arguments to arbitrary SU(4) gates requires each step of the argument to be *robust*, in the sense that small errors in each step do not accumulate significantly. In particular, we can only approximately disentangle the state using the procedure in (9), and learning the remaining 1D states poses new technical challenges as they are no longer pure. These issues are addressed in Section 8.4, which leads to a robust version of the above result; see Claim 1 of Theorem 9.

3 Preliminaries

Let $\text{stab}_1 = \{|0\rangle, |1\rangle, |+\rangle, |-\rangle, |y+\rangle, |y-\rangle\}$ be the set of single-qubit stabilizer states. Given an n-qubit unitary U, we use the Catholic letter $\mathcal{U}$ to denote the corresponding CPTP map $\mathcal{U}(X) = UXU^\dagger$. We denote $\mathcal{I}$ as the identity CPTP map. Given a Pauli operator $P \in \{X, Y, Z\}$, we consider P_i to be a multi-qubit operator that is equal to the tensor product of P on the i-th qubit and identity on the rest of the qubits. We also consider the following definitions.

Definition 1 (Reduced channel)**.** *Given $n > 0$, $i \in \{1, \ldots, n\}$, and an n-qubit CPTP map $\mathcal{C}$. The reduced channel $\mathcal{E}^{\mathcal{C}}_{\neq i}$ of the CPTP map $\mathcal{C}$ with the i-th qubit removed is*

$$\mathcal{E}^{\mathcal{C}}_{\neq i}(\rho_{\neq i}) = \text{Tr}_i\left(\mathcal{C}\left(\frac{I^{(i)}}{2} \otimes \rho_{\neq i}\right)\right), \tag{10}$$

where $\rho_{\neq i}$ is a density matrix on all except the i-th qubit, $I^{(i)}$ is the identity on the i-th qubit, and Tr_i is the partial trace over the i-th qubit. For $k \in \{0, 1, \ldots, n\}$, we define

$$\mathcal{E}^{\mathcal{C}}_{>k}(\rho_{>k}) = \text{Tr}_{\leq k}\left(\mathcal{C}\left(\frac{I^{(1,\ldots,k)}}{2^k} \otimes \rho_{>k}\right)\right), \tag{11}$$

where $\rho_{>k}$ is a density matrix on all except the first k qubits, $I^{(1,\ldots,k)}$ is the identity on the first k qubits, and $\text{Tr}_{\leq k}$ is the partial trace over the first k qubits. Given a subset of qubits $S \subseteq \{1, \ldots, n\}$, we define

$$\mathcal{E}^{\mathcal{C}}_{S}(\cdot) = \text{Tr}_{\notin S}\left(\mathcal{C}\left(\frac{I^{(\notin S)}}{2^{n-|S|}} \otimes (\cdot)\right)\right), \tag{12}$$

where $I^{(\notin S)}$ is the identity on qubits not in S and $\text{Tr}_{\notin S}$ is the partial trace over qubits not in S.

Definition 2 (Fidelity). *Given two quantum states ρ, σ. The fidelity $\mathcal{F}(\rho,\sigma) \in [0,1]$ between the two states is defined as $\mathrm{Tr}\big(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\big)^2$. If $\sigma = |\psi\rangle\langle\psi|$, then $\mathcal{F}(\rho,\sigma) = \langle\psi|\,\rho\,|\psi\rangle$.*

Fact 1 (Properties of fidelity [99]). *The function $1 - F(\rho,\sigma)$ satisfies*

$$1 - F(\rho,\sigma) = 1 - F(\sigma,\rho) \qquad \textit{(symmetric)}; \tag{13}$$
$$1 - F(\rho,\sigma) \geq 0 \qquad \textit{(nonnegative)}; \tag{14}$$
$$1 - F(\rho,\sigma) = 0 \iff \rho = \sigma \qquad \textit{(identity of indiscernible)}. \tag{15}$$

But $1 - F$ does not satisfy triangle inequality. In contrast, $\Theta(\rho,\sigma) := \arcsin\Big(\sqrt{1 - F(\rho,\sigma)}\Big) \in [0,\pi/2]$ is symmetric, nonnegative, and satisfies identity of indiscernible and triangle inequality,

$$\Theta(\rho,\sigma) \leq \Theta(\rho,\tau) + \Theta(\tau,\sigma). \tag{16}$$

Hence, $\theta(\rho,\sigma)$ is a metric (known as the Fubini-Study metric), but $1 - F(\rho,\sigma)$ is not. In addition to the metric properties, we also have

$$1 - F(\psi,\rho) \leq \frac{1}{2}\|\psi - \rho\|_{\mathrm{tr}}, \tag{17}$$

for any state ρ and any pure state ψ, where $\|\cdot\|_{\mathrm{tr}}$ is the trace norm. Also, the fidelity is monotonic increasing under CPTP maps,

$$F(\mathcal{E}(\rho),\mathcal{E}(\sigma)) \geq F(\rho,\sigma), \tag{18}$$

for any CPTP map $\mathcal{E}$ and any state ρ, σ.

Definition 3 (Average-case distance). *Given two n-qubit CPTP maps $\mathcal{E}_1, \mathcal{E}_2$. The average-case distance $\mathcal{D}_{\mathrm{ave}}(\mathcal{E}_1,\mathcal{E}_2)$ between the two CPTP maps is defined as*

$$\mathop{\mathbb{E}}_{|\psi\rangle:\mathrm{Unif}} \big[1 - \mathcal{F}(\mathcal{E}_1(|\psi\rangle\langle\psi|), \mathcal{E}_2(|\psi\rangle\langle\psi|))\big], \tag{19}$$

where $\mathbb{E}_{|\psi\rangle:\mathrm{Unif}}$ considers averaging under the uniform measure over pure states.

Fact 2 (Haar average for average-case distance [96]). *Given an n-qubit CPTP map $\mathcal{E}$ and an n-qubit unitary U. We have the following identity,*

$$\mathcal{D}_{\mathrm{ave}}(\mathcal{E},\mathcal{U}) = \frac{2^n}{2^n+1}\left(1 - \frac{1}{4^n}\sum_{i,j}\langle i|\,\mathcal{E}\left(U^\dagger\,|i\rangle\langle j|\,U\right)|j\rangle\right), \tag{20}$$

after averaging over the uniform measure over pure states.

Proposition 1 (Normalized Frobenius norm). *Given two n-qubit unitaries U_1, U_2. We have*

$$\frac{1}{3}\min_{\phi\in\mathbb{R}}\frac{\left\|e^{i\phi}U_1 - U_2\right\|_F^2}{2^n} \leq \mathcal{D}_{\mathrm{ave}}(\mathcal{U}_1,\mathcal{U}_2) \leq \min_{\phi\in\mathbb{R}}\frac{\left\|e^{i\phi}U_1 - U_2\right\|_F^2}{2^n}, \tag{21}$$

where $\|X\|_F = \sqrt{\mathrm{Tr}(X^\dagger X)}$ is the Frobenius norm of X.

Proof. From [96], the average-case distance (also known as the average gate fidelity) satisfies

$$\mathcal{D}_{\text{ave}}(\mathcal{U}_1, \mathcal{U}_2) = \frac{2^n}{2^n+1}\left(1 - \frac{1}{4^n}\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|^2\right). \tag{22}$$

Expanding the definition of Frobenius norm, we have

$$\min_{\phi\in\mathbb{R}} \frac{\left\|e^{i\phi}U_1 - U_2\right\|_F^2}{2^n} = 2\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right). \tag{23}$$

Recall that

$$0 \le \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n} \le 1. \tag{24}$$

Hence, we have

$$\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right) \le \left(1 + \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right)\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right) \le 2\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right). \tag{25}$$

This immediately implies that

$$\frac{2}{3}\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right) \le \frac{2^n}{2^n+1}\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|^2}{4^n}\right) \le 2\left(1 - \frac{\left|\text{Tr}\left(U_1^\dagger U_2\right)\right|}{2^n}\right) \tag{26}$$

which is equivalent to

$$\frac{1}{3}\min_{\phi\in\mathbb{R}} \frac{\left\|e^{i\phi}U_1 - U_2\right\|_F^2}{2^n} \le \mathcal{D}_{\text{ave}}(\mathcal{U}_1, \mathcal{U}_2) \le \min_{\phi\in\mathbb{R}} \frac{\left\|e^{i\phi}U_1 - U_2\right\|_F^2}{2^n}. \tag{27}$$

This concludes the proof. □

Definition 4 (Worse-case distance / diamond distance)**.** *Given two n-qubit CPTP maps $\mathcal{E}_1, \mathcal{E}_2$. The worst-case distance $\mathcal{D}_\diamond(\mathcal{E}_1, \mathcal{E}_2)$ between the two CPTP maps is defined as*

$$\frac{1}{2}\max_\rho \left\|(\mathcal{E}_1 \otimes \mathcal{I})(\rho) - (\mathcal{E}_2 \otimes \mathcal{I})(\rho)\right\|_1 \triangleq \frac{1}{2}\left\|\mathcal{E}_1 - \mathcal{E}_2\right\|_\diamond, \tag{28}$$

where ρ is maximized over $2n$-qubit states and $\mathcal{I}^{(>n)}$ is an identity map acting on the n qubits. $\mathcal{D}_\diamond(\mathcal{E}_1, \mathcal{E}_2)$ is also known as diamond distance and $\|\cdot\|_\diamond$ is the diamond norm.

Fact 3 (Diamond distance for unitaries; Prop. 1.6 of [100])**.** *For any two unitaries U_1, U_2, we have*

$$\min_{\phi\in\mathbb{R}} \left\|e^{i\phi}U_1 - U_2\right\|_\infty \le \left\|\mathcal{U}_1 - \mathcal{U}_2\right\|_\diamond \le 2\min_{\phi\in\mathbb{R}} \left\|e^{i\phi}U_1 - U_2\right\|_\infty. \tag{29}$$

Fact 4 (Exact unitary synthesis; see e.g. [101, 102])**.** *Given any unitary U acting on k qubits, there is an algorithm that outputs a circuit (acting on k qubits) consisting of at most 4^k two-qubit gates, which exactly implements the unitary U, in time $2^{O(k)}$.*

Corollary 1 (Exact unitary synthesis in geometrically-local circuit)**.** *Given any unitary U acting on k qubits and a connected graph G over k qubits, there is an algorithm that outputs a geometrically-local circuit (acting on k qubits and consists only of gates between connected qubits) consisting of at most $2k4^k$ two-qubit gates, which exactly implements the unitary U, in time $2^{O(k)}$.*

Proof. For each two-qubit gate in the original synthesis protocol, which may not be geometrically-local under the connectivity graph G, we consider at most $k-1$ swap gates to move one of the qubits from the original location to a location next to the other qubit, apply the two-qubit gate, then perform at most $k-1$ swap gates to move the qubit back to the original location. □

4 Approximate local identity

A central concept that we will use to define local inversion for representing n-qubit unitaries is the ε-*approximate local identity*. In this section, we provide the properties for understanding the concept of approximate local identity. In particular, we will consider a strong and a weak form of local identity in Section 4.1 and 4.2. In each section, we state the definition, show how to characterize if a unitary map forms a strong/weak ε-approximate local identity, and prove how local identity relates to global identity.

4.1 Strong ε-approximate local identity

We begin by looking at a strong form of approximate local identity. The idea is that the action of the n-qubit unitary U on the i-th qubit is close to the identity map, while the action on the other qubits is close to the reduced channel of U with the i-th qubit removed (feed in a maximally mixed state on qubit i and trace out qubit i at the end). Recall Definition 1 of reduced channel,

$$\mathcal{E}^{\mathcal{U}}_{\neq i}(\rho_{\neq i}) = \mathrm{Tr}_i\left(\mathcal{U}\left(\frac{I^{(i)}}{2}\otimes\rho_{\neq i}\right)\right), \tag{30}$$

where $\rho_{\neq i}$ is a density matrix on all except the i-th qubit, $I^{(i)}$ is the identity on the i-th qubit, and Tr_i is the partial trace over the i-th qubit.

Definition 5 (Strong ε-approximate local identity)**.** *Given $n>0, \varepsilon\geq 0$, and $i\in\{1,\ldots,n\}$. An n-qubit unitary U is a strong ε-approximate local identity on the i-th qubit if*

$$\mathcal{D}_\diamond\left(\mathcal{U}, \mathcal{I}^{(i)}\otimes\mathcal{E}^{\mathcal{U}}_{\neq i}\right)\leq\varepsilon, \tag{31}$$

where $\mathcal{I}^{(i)}\otimes\mathcal{E}^{\mathcal{U}}_{\neq i}$ is an n-qubit CPTP map that acts as identity on the i-th qubit.

While diamond distances are typically hard to characterize, the strong ε-approximate local identity can be characterized up to a constant factor by studying the Heisenberg evolution of single-qubit Pauli observables under the n-qubit unitary U. Hence, in order to check if an n-qubit unitary U strong approximate local identity on the i-th qubit, all we need to check is whether the three Pauli observables X_i, Y_i, Z_i remains approximately unchanged after Heisenberg evolution under U.

Lemma 1 (Characterization of strong ε-approximate local identity)**.** *Given $n>0$, $\varepsilon\geq 0$, and an n-qubit unitary $\mathcal{U}$. If $\mathcal{U}$ is a strong ε-approximate local identity on the i-th qubit, then*

$$\frac{1}{2}\left\|U^\dagger P_i U - P_i\right\|_\infty \leq \varepsilon, \forall P\in\{X,Y,Z\}, \tag{32}$$

where P_i is the Pauli operator P acting only on qubit i, and $U^\dagger P_i U$ is the Heisenberg evolution of P_i under U. Furthermore, if the following holds,

$$\frac{1}{2}\sum_{P\in\{X,Y,Z\}}\left\|U^\dagger P_i U - P_i\right\|_\infty \leq \varepsilon, \tag{33}$$

then $\mathcal{U}$ is a strong ε-approximate local identity on the i-th qubit.

Proof. We start by showing the first claim. Consider any n-qubit pure state $|\psi\rangle$. We have

$$\left\|U^\dagger P_i U - P_i\right\|_\infty = \max_{|\psi\rangle}\left|\langle\psi|\left(U^\dagger P_i U - P_i\right)|\psi\rangle\right|. \tag{34}$$

By the definition of CPTP maps, we have

$$\langle\psi|U^\dagger P_i U|\psi\rangle = \mathrm{Tr}\left(P_i\mathcal{U}\left(|\psi\rangle\langle\psi|\right)\right). \tag{35}$$

From the definition of diamond distance and of strong ε-approximate local identity on the i-th qubit, we have the following inequality,

$$\frac{1}{2}\left|\mathrm{Tr}\left(P_i\mathcal{U}\left(|\psi\rangle\langle\psi|\right)\right) - \mathrm{Tr}\left(P_i\left(\mathcal{I}^{(i)}\otimes\mathcal{E}^{\mathcal{U}}_{\neq i}\right)\left(|\psi\rangle\langle\psi|\right)\right)\right| \leq \varepsilon. \tag{36}$$

By the definition of a CPTP map, we have

$$\mathrm{Tr}_{\neq i}\left(\mathcal{E}^{\mathcal{U}}_{\neq i}(\rho)\right) = \rho \tag{37}$$

for any quantum state ρ, where $\mathrm{Tr}_{\neq i}$ traces out all qubits except for qubit i. Hence, we have $\mathrm{Tr}\left(P_i\left(\mathcal{I}^{(i)}\otimes\mathcal{E}^{\mathcal{U}}_{\neq i}\right)\left(|\psi\rangle\langle\psi|\right)\right) = \mathrm{Tr}(P_i|\psi\rangle\langle\psi|)$. Together, we obtain the first claim.

The second claim uses the following equality defined over an $n+1$-qubit system,

$$\frac{1}{2}\left(I_{n+1} + \sum_{P\in\{X,Y,Z\}} P_i\otimes P\right) = S_{i,n+1}, \tag{38}$$

where I_{n+1} is an $n+1$-qubit identity, P_i is an n-qubit unitary that acts as the Pauli operator P on the i-th qubit, and $S_{i,n+1}$ is the swap operator between qubit i in the first n qubits and the last qubit (qubit $n+1$). We interpret the error in the Heisenberg-evolved single-qubit Pauli observables as an error in commuting the Pauli observable P_i and the n-qubit unitary U,

$$\left\|U^\dagger P_i U - P_i\right\|_\infty = \left\|P_i U - U P_i\right\|_\infty. \tag{39}$$

From this interpretation, we have the following inequalities,

$$\begin{aligned}
\left\|S_{i,n+1}(U\otimes I) - (U\otimes I)S_{i,n+1}\right\|_\infty &\leq \frac{1}{2}\sum_{P\in\{X,Y,Z\}}\left\|(P_i\otimes P)(U\otimes I) - (U\otimes I)(P_i\otimes P)\right\|_\infty &(40)\\
&\leq \frac{1}{2}\sum_{P\in\{X,Y,Z\}}\left\|(P_iU - UP_i)\otimes P\right\|_\infty &(41)\\
&= \frac{1}{2}\sum_{P\in\{X,Y,Z\}}\left\|P_iU - UP_i\right\|_\infty \leq \varepsilon. &(42)
\end{aligned}$$

The above inequality can be easily generalized to any of the following,

$$\|S_{i,j}(U \otimes I_m) - (U \otimes I_m)S_{i,j}\|_\infty \leq \varepsilon, \tag{43}$$

where $m \geq 1$, $n+1 \leq j \leq n+m$, and I_m is the identity operator on m qubits. Recall the formal definition diamond distance from Definition 4,

$$\mathcal{D}_\diamond(\mathcal{E}_1, \mathcal{E}_2) = \frac{1}{2}\max_\rho \|(\mathcal{E}_1 \otimes \mathcal{I}_n)(\rho) - (\mathcal{E}_2 \otimes \mathcal{I}_n)(\rho)\|_1, \tag{44}$$

where ρ is a density matrix over $2n$ qubits, and $\mathcal{I}_n$ is the identity map over n qubits. From Fact 3, for any two unitaries U_1, U_2, we have $\|\mathcal{U}_1 - \mathcal{U}_2\|_\diamond \leq 2\|U_1 - U_2\|_\infty$. We obtain the following from Eq. (43),

$$\left\|\mathcal{S}_{i,j}(\mathcal{U} \otimes I_m) - (\mathcal{U} \otimes I_m)\mathcal{S}_{i,j}\right)\right\|_\diamond \leq 2\|S_{i,j}(U \otimes I_m) - (U \otimes I_m)S_{i,j}\|_\infty \leq 2\varepsilon. \tag{45}$$

The strong ε-approximate local identity considers

$$\mathcal{D}_\diamond\left(\mathcal{U}, \mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i}\right) = \frac{1}{2}\max_\rho \left\|(\mathcal{U} \otimes \mathcal{I}_n)(\rho) - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_n)(\rho)\right\|_1. \tag{46}$$

We add one more qubit to form $2n+1$ qubits. The additional qubit begins in a maximally mixed state $I/2$, stays in $I/2$, and is traced out at the end. Let us now consider the following series of analysis,

$$\begin{aligned}
&\left\|(\mathcal{U} \otimes \mathcal{I}_n)(\rho) - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_n)(\rho)\right\|_1 && (47)\\
&= \left\|\mathrm{Tr}_{2n+1}\left[(\mathcal{U} \otimes \mathcal{I}_{n+1})(\rho \otimes (I/2))\right] - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_n)(\rho)\right\|_1 && (48)\\
&= \left\|\mathrm{Tr}_i\left[(\mathcal{S}_{i,2n+1} \circ (\mathcal{U} \otimes \mathcal{I}_{n+1}))\,(\rho \otimes (I/2))\right] - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_{n+1})(\rho \otimes (I/2))\right\|_1 && (49)\\
&\leq \left\|\mathrm{Tr}_i\left[((\mathcal{U} \otimes \mathcal{I}_{n+1}) \circ \mathcal{S}_{i,2n+1})\,(\rho \otimes (I/2))\right] - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_{n+1})(\rho \otimes (I/2))\right\|_1 + 2\varepsilon && (50)\\
&= \left\|(\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_{n+1})(\rho \otimes (I/2)) - (\mathcal{I}^{(i)} \otimes \mathcal{E}^{\mathcal{U}}_{\neq i} \otimes \mathcal{I}_{n+1})(\rho \otimes (I/2))\right\|_1 + 2\varepsilon = 2\varepsilon. && (51)
\end{aligned}$$

The only inequality above uses Eq. (45). We have proved the claim. □

The following two lemmas give the relationships between global and local identity checks. The basic idea is to check whether a map is close to identity by checking whether the map forms approximate local identities on all the n qubits. If the map is far from identity, then the map is not an approximate local identity for some qubits. If the map is an approximate local identity for all qubits, then the map is close to the identity.

Lemma 2 (Global non-identity check from local non-identity checks)**.** *Given an integer $n > 0$ and an n-qubit unitary U. If there exists $\varepsilon > 0$ and $i \in \{1, \ldots, n\}$, such that $\mathcal{U}$ is not a strong ε-approximate local identity on the i-th qubit, then $\|\mathcal{U} - \mathcal{I}\|_\diamond \geq \varepsilon/2$.*

Lemma 3 (Global identity check from local identity checks)**.** *Given an integer $n > 0$ and an n-qubit unitary U. If there exists $\varepsilon_1, \ldots, \varepsilon_n > 0$, such that $\mathcal{U}$ is a strong ε_i-approximate local identity on the i-th qubit for all $i \in \{1, \ldots, n\}$, then $\|\mathcal{U} - \mathcal{I}\|_\diamond \leq 3\sum_{i=1}^n \varepsilon_i$.*

We give proofs of these two lemmas at the end of this subsection. Lemma 2 is proven by contradiction. To prove Lemma 3, we consider a stabilizer decomposition for a single qubit.

Proposition 2 (Single-qubit stabilizer decomposition)**.** *Given an integer $n > 0$ and an n-qubit density matrix ρ. For any $S \subseteq \{1, \ldots, n\}$, ρ can be written as a linear combination of $R = 10^{|S|}$ n-qubit density matrices $\rho_1, \ldots, \rho_R$, $\rho = \sum_{r=1}^{R} \alpha_r \rho_r$, where $\alpha_r \in \mathbb{R}$ and ρ_r is a density matrix that satisfies*

$$\rho_r = \bigotimes_{j \in S} |s_j\rangle\langle s_j| \otimes \mathrm{Tr}_S(\rho_r), \tag{52}$$

for some $|s_j\rangle \in \mathrm{stab}_1$. We also have $\sum_{r=1}^{R} \alpha_r = 1$ and $\sum_{r=1}^{R} |\alpha_r| = 3^{|S|}$.

Proof. Given an integer $i \in \{1, \ldots, n\}$, consider the following linear map $\mathcal{M}_i$ which equals to the identity channel on i-th qubit,

$$\begin{aligned}
\mathcal{M}_i(\rho) := {} & |0\rangle\langle 0|_i \otimes \langle 0| \rho |0\rangle_i + |1\rangle\langle 1|_i \otimes \langle 1| \rho |1\rangle_i \\
& + \frac{1}{2} |+\rangle\langle +|_i \otimes \langle +| \rho |+\rangle_i - \frac{1}{2} |+\rangle\langle +|_i \otimes \langle -| \rho |-\rangle_i \\
& - \frac{1}{2} |-\rangle\langle -|_i \otimes \langle +| \rho |+\rangle_i + \frac{1}{2} |-\rangle\langle -|_i \otimes \langle -| \rho |-\rangle_i \qquad (53) \\
& + \frac{1}{2} |y+\rangle\langle y+|_i \otimes \langle y+| \rho |y+\rangle_i - \frac{1}{2} |y+\rangle\langle y+|_i \otimes \langle y-| \rho |y-\rangle_i \\
& - \frac{1}{2} |y-\rangle\langle y-|_i \otimes \langle y+| \rho |y+\rangle_i + \frac{1}{2} |y-\rangle\langle y-|_i \otimes \langle y-| \rho |y-\rangle_i , \\
= {} & \sum_{r=1}^{10} b_r |s_r\rangle\langle s_r|_i \otimes \langle s'_r| \rho |s'_r\rangle_i . \qquad (54)
\end{aligned}$$

where $|s\rangle\langle s|_i$ is a single-qubit stabilizer state on the i-th qubit, $\langle s| \rho |s\rangle_i$ is a partial inner product on the i-th qubit, s_r, s'_r, b_r takes on the corresponding values in stab_1, stab_1, $\{1, 1/2, -1/2\}$, respectively. The fact that $\mathcal{M}_i$ equals to the identity CPTP map $\mathcal{I}$ is because of the following identity

$$\rho = \sum_{P \in \{I,X,Y,Z\}} \mathrm{Tr}_i(P_i \rho) \otimes \frac{P_i}{2}, \tag{55}$$

where P_i acts on the i-th qubit, and Eq. (53) follows by further decomposing the Pauli operators into their eigenstates.

Without loss of generality, we consider $k = |S|$ and $S = \{1, \ldots, k\}$. The identity $\rho = (\circ_{i \in S} \mathcal{M}_i)(\rho)$ gives rise to the equality

$$\rho = \sum_{r_1=1}^{10} \cdots \sum_{r_k=1}^{10} \left(\prod_{i=1}^{k} b_{r_i} \right) |s_{r_1}, \ldots, s_{r_k}\rangle\langle s_{r_1}, \ldots, s_{r_k}| \otimes \langle s'_{r_1}, \ldots, s'_{r_k}| \rho |s'_{r_1}, \ldots, s'_{r_k}\rangle . \tag{56}$$

We define $r = \sum_{i=1}^{k} 10^{i-1} r_i$, $R = 10^k$, $Z_r = \mathrm{Tr}\big(\langle s'_{r_1}, \ldots, s'_{r_k}| \rho |s'_{r_1}, \ldots, s'_{r_k}\rangle\big) \geq 0$, and

$$\rho_r = \begin{cases} |s_{r_1}, \ldots, s_{r_k}\rangle\langle s_{r_1}, \ldots, s_{r_k}| \otimes \frac{\langle s'_{r_1}, \ldots, s'_{r_k}| \rho |s'_{r_1}, \ldots, s'_{r_k}\rangle}{Z_r} & \text{if } Z_r > 0, \\ |s_{r_1}, \ldots, s_{r_k}\rangle\langle s_{r_1}, \ldots, s_{r_k}| \otimes \frac{I}{2^{n-k}} & \text{if } Z_r = 0, \end{cases} \tag{57}$$

and $\alpha_r = Z_r \prod_{i=1}^{k} b_{r_i}$. It is not hard to check that $\sum_r |\alpha_r| = 3^k$. Together, we have the single-qubit stabilizer decomposition $\rho = \sum_{r=1}^{R} \alpha_r \rho_r$. □

Proof of Lemma 2. We consider proof by contradiction. Assume $\|\mathcal{U}-\mathcal{I}\|_\diamond < \varepsilon/2$. For any integer $m \geq 0$, for any state $|s\rangle_i \in \text{stab}_1$ on the i-th qubit, and for any $(n-1+m)$-qubit density matrix ρ,

$$\left\|(\mathcal{U}\otimes\mathcal{I}^{(>n)})\left(|s\rangle\langle s|_i\otimes\rho\right)-|s\rangle\langle s|_i\otimes(\mathcal{E}^U_{\neq i}\otimes\mathcal{I}^{(>n)})(\rho)\right\|_1 \tag{58}$$

$$\leq \left\|(\mathcal{U}\otimes\mathcal{I}^{(>n)})\left(|s\rangle\langle s|_i\otimes\rho\right)-|s\rangle\langle s|_i\otimes\rho\right\|_1+\left\||s\rangle\langle s|_i\otimes\rho-|s\rangle\langle s|_i\otimes(\mathcal{E}^U_{\neq i}\otimes\mathcal{I}^{(>n)})(\rho)\right\|_1 \tag{59}$$

$$\leq \|\mathcal{U}-\mathcal{I}\|_\diamond+\|\mathcal{U}-\mathcal{I}\|_\diamond<\varepsilon. \tag{60}$$

The first inequality follows from putting in $|s\rangle\langle s|_i\otimes\rho$ and using triangle inequality. The second inequality follows from the definition of diamond distance, the identity

$$\left\||s\rangle\langle s|_i\otimes\rho-|s\rangle\langle s|_i\otimes(\mathcal{E}^U_{\neq i}\otimes\mathcal{I}^{(>n)})(\rho)\right\|_1 \tag{61}$$

$$=\left\||s\rangle\langle s|_i\otimes\text{Tr}_i\left(\frac{I^{(i)}}{2}\otimes\rho\right)-|s\rangle\langle s|_i\otimes\text{Tr}_i\left(\left(\mathcal{U}\otimes\mathcal{I}^{(>n)}\right)\left(\frac{I^{(i)}}{2}\otimes\rho\right)\right)\right\|_1, \tag{62}$$

and the two facts: $\|\rho_A\otimes\rho_B-\rho_A\otimes\rho_C\|_1=\|\rho_B-\rho_C\|_1, \|\text{Tr}_i(\rho_A)\|_1\leq\|\text{Tr}(\rho_A)\|_1$ for any density matrix ρ_A,ρ_B,ρ_C. The above derivation shows that U is an ε-approximate local identity on the i-th qubit, which is a contradiction. Therefore, $\|\mathcal{U}-\mathcal{I}\|_\diamond\geq\varepsilon/2$. □

Proof of Lemma 3. From Theorem 3.55 in [103], we have

$$\|\mathcal{U}-\mathcal{I}\|_\diamond=\left\|U|\psi\rangle\langle\psi|U^\dagger-|\psi\rangle\langle\psi|\right\|_1 \tag{63}$$

for some n-qubit state $|\psi\rangle$. Let $\mathcal{I}^{(\leq k)}$ be the identity CPTP map acting on the first k qubit. We use a telescoping sum of the form,

$$U|\psi\rangle\langle\psi|U^\dagger-|\psi\rangle\langle\psi|=\sum_{k=0}^{n-1}\left[\left(\mathcal{I}^{(\leq k)}\otimes\mathcal{E}^U_{>k}\right)(|\psi\rangle\langle\psi|)-\left(\mathcal{I}^{(\leq k+1)}\otimes\mathcal{E}^U_{>k+1}\right)(|\psi\rangle\langle\psi|)\right]. \tag{64}$$

By triangle inequality, we obtain

$$\|\mathcal{U}-\mathcal{I}\|_\diamond\leq\sum_{k=0}^{n-1}\left\|\left(\mathcal{I}^{(\leq k)}\otimes\mathcal{E}^U_{>k}\right)(|\psi\rangle\langle\psi|)-\left(\mathcal{I}^{(\leq k+1)}\otimes\mathcal{E}^U_{>k+1}\right)(|\psi\rangle\langle\psi|)\right\|_1. \tag{65}$$

In the next step, we will bound each term in the above telescoping sum.

To bound the term corresponding to $k\in\{0,\ldots,n-1\}$ in Eq. (65), we consider an $(k+(n-k)+k)$-qubit density matrix $\rho^{(k)}$. The first k qubits of $\rho^{(k)}$ is the maximally mixed state $\frac{I^{(1,\ldots,k)}}{2^k}$. The next $(n-k)$ qubits of $\rho^{(k)}$ corresponds to all except the first k qubits in $|\psi\rangle\langle\psi|$. The last k qubits of $\rho^{(k)}$ corresponds to the first k qubits in $|\psi\rangle\langle\psi|$. Under this definition of $\rho^{(k)}$, we have

$$\left\|\left(\mathcal{I}^{(\leq k)}\otimes\mathcal{E}^U_{>k}\right)(|\psi\rangle\langle\psi|)-\left(\mathcal{I}^{(\leq k+1)}\otimes\mathcal{E}^U_{>k+1}\right)(|\psi\rangle\langle\psi|)\right\|_1 \tag{66}$$

$$=\left\|\left(\mathcal{U}\otimes\mathcal{I}^{(>n)}\right)(\rho^{(k)})-\left(\mathcal{I}^{(k+1)}\otimes\mathcal{E}^U_{\neq k+1}\otimes\mathcal{I}^{(>n)}\right)(\rho^{(k)})\right\|_1, \tag{67}$$

where $\left(\mathcal{I}^{(k+1)}\otimes\mathcal{E}^U_{\neq k+1}\otimes\mathcal{I}^{(>n)}\right)(\rho^{(k)})$ is the output state after applying the $(n-1)$-qubit CPTP map $\mathcal{E}^U_{\neq k+1}$ to the first n qubits except the $(k+1)$-th qubit of $\rho^{(k)}$. We now use the single-qubit stabilizer decomposition with $S=\{k+1\}$ given in Prop. 2 to obtain $\rho^{(k)}=\sum_{r=1}^{10}\alpha_r\rho_r^{(k)}$ with

$\sum_r |\alpha_r| = 3$ and the reduced density matrix of $\rho_r^{(k)}$ on the $(k+1)$-th qubit is a single-qubit stabilizer state. We can now bound each term by

$$\left\| \left(\mathcal{U} \otimes \mathcal{I}^{(>n)} \right) (\rho^{(k)}) - \left(\mathcal{I}^{(k+1)} \otimes \mathcal{E}^U_{\neq k+1} \otimes \mathcal{I}^{(>n)} \right) (\rho^{(k)}) \right\|_1 \tag{68}$$

$$\leq \sum_{r=1}^{10} |\alpha_r| \left\| \left(\mathcal{U} \otimes \mathcal{I}^{(>n)} \right) (\rho_r^{(k)}) - \left(\mathcal{I}^{(k+1)} \otimes \mathcal{E}^U_{\neq k+1} \otimes \mathcal{I}^{(>n)} \right) (\rho_r^{(k)}) \right\|_1 \tag{69}$$

$$\leq \sum_{r=1}^{10} |\alpha_r| \varepsilon_{k+1} = 3\varepsilon_{k+1}. \tag{70}$$

The first line is the triangle inequality. The second line uses the assumption that U is an ε_{k+1}-approximate local identity on the $(k+1)$-th qubit. Combining Eq. (65), Eq. (67), Eq. (70),

$$\|\mathcal{U} - \mathcal{I}\|_\diamond \leq 3 \sum_{k=0}^{n-1} \varepsilon_{k+1}, \tag{71}$$

which establishes the stated result. □

4.2 Weak ε-approximate local identity

We next look at another definition of approximate local identity: the reduced channel of U on the i-th qubit is close to the identity map. This definition is very easy to check but only guarantees that the unitary U is close to the identity in the average-case distance (instead of the worst-case distance, i.e., the diamond distance). Hence, we will refer to this as the weak ε-approximate local identity. Recall Definition 1 of reduced channel,

$$\mathcal{E}_i^{\mathcal{U}}(\rho_i) = \mathrm{Tr}_{\neq i} \left(U \left(\frac{I^{(\neq i)}}{2^{n-1}} \otimes \rho_i \right) U^\dagger \right), \tag{72}$$

where ρ_i is a density matrix on the i-th qubit, $I^{(\neq i)}$ is the identity on all except the i-th qubit, and $\mathrm{Tr}_{\neq i}$ is the partial trace over all except the i-th qubit.

Definition 6 (Weak ε-approximate local identity; unitary version)**.** *Given $n > 0, \varepsilon \geq 0$, and $i \in \{1, \ldots, n\}$. An n-qubit unitary U is a weak ε-approximate local identity on the i-th qubit if*

$$\mathcal{D}_{\text{ave}} \left(\mathcal{E}_i^{\mathcal{U}}, \mathcal{I} \right) \leq \varepsilon, \tag{73}$$

where $\mathcal{I}$ is a 1-qubit CPTP map that acts as an identity.

In the literature of quantum junta learning [104], one defines the influence of a qubit i in an n-qubit unitary $U = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \alpha_P P$, where $\alpha_P \in \mathbb{C}$ to be

$$\sum_{\substack{P \in \{I,X,Y,Z\}^{\otimes n} \\ P_i \neq I}} |\alpha_P|^2 . \tag{74}$$

The following lemma shows that weak approximate local identity is equivalent to low influence.

Lemma 4 (Characterization of weak ε-approximate local identity)**.** *Given $n > 0$, $\varepsilon \geq 0$, and an n-qubit unitary U. Consider the Pauli representation of $U = \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} \alpha_P P$, where $\alpha_P \in \mathbb{C}$. $\mathcal{U}$ is a weak ε-approximate local identity on the i-th qubit if and only if*

$$\sum_{\substack{P \in \{I,X,Y,Z\}^{\otimes n} \\ P_i \neq I}} |\alpha_P|^2 \leq \frac{3}{2}\varepsilon. \tag{75}$$

From the definition of influence in quantum junta learning [104], we have qubit i has influence bounded above by 1.5ε in the unitary U.

Proof. From the definition of the reduced channel, we have

$$\mathcal{E}_i^{\mathcal{U}}(\rho_i) = \sum_{s_1, s_2 \in \{I,X,Y,Z\}} \left(\sum_{\substack{P,Q \in \{I,X,Y,Z\}^{\otimes n} \\ P_i = s_1, Q_i = s_2, P_{\neq i} = Q_{\neq i}}} \alpha_P^* \alpha_Q \right) s_1 \rho_i s_2, \tag{76}$$

where $P_{\neq i}, Q_{\neq i}$ is an $(n-1)$-qubit Pauli observable equal to P, Q with qubit i removed. From Fact 2 characterizing the average-case distance $\mathcal{D}_{\text{ave}}$, we have

$$\mathcal{D}_{\text{ave}}\left(\mathcal{E}_i^{\mathcal{U}}, \mathcal{I}\right) = \frac{2}{3}\left(1 - \sum_{\substack{P,Q \in \{I,X,Y,Z\}^{\otimes n} \\ P_i = I, Q_i = I, P_{\neq i} = Q_{\neq i}}} \alpha_P^* \alpha_Q\right) = \frac{2}{3}\left(1 - \sum_{\substack{P \in \{I,X,Y,Z\}^{\otimes n} \\ P_i = I}} |\alpha_P|^2\right). \tag{77}$$

Furthermore, we note that $\operatorname{Tr}\left(U^\dagger U\right) = 2^n = 2^n \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} |\alpha_P|^2$. Hence, we have

$$1 - \sum_{\substack{P \in \{I,X,Y,Z\}^{\otimes n} \\ P_i = I}} |\alpha_P|^2 = \sum_{\substack{P \in \{I,X,Y,Z\}^{\otimes n} \\ P_i \neq I}} |\alpha_P|^2. \tag{78}$$

The lemma follows from the two identities given above. □

Weak ε-approximate local identity naturally generalizes to any quantum process (channel) by using the definition of reduced channels for channels. The formal definition is given below.

Definition 7 (Weak ε-approximate local identity; channel version)**.** *Given $n > 0, \varepsilon \geq 0$, and $i \in \{1, \ldots, n\}$. An n-qubit CPTP map $\mathcal{C}$ is a weak ε-approximate local identity on the i-th qubit if*

$$\mathcal{D}_{\text{ave}}\left(\mathcal{E}_i^{\mathcal{C}}, \mathcal{I}\right) \leq \varepsilon, \tag{79}$$

where $\mathcal{I}$ is a 1-qubit CPTP map that acts as an identity.

The following two lemmas give the relationships between global and local identity checks. The basic idea is to check whether a map is close to identity by checking whether the map forms approximate local identities on all the n qubits.

Lemma 5 (Global non-identity check from local non-identity checks)**.** *Given an integer $n > 0$ and an n-qubit CPTP map $\mathcal{C}$. If there exists $\varepsilon > 0$ and $i \in \{1, \ldots, n\}$, such that $\mathcal{C}$ is not a weak ε-approximate local identity on the i-th qubit, then $\mathcal{D}_{\text{ave}}(\mathcal{C}, \mathcal{I}) \geq \varepsilon$.*

Lemma 6 (Global identity check from local identity checks)**.** *Given an integer $n > 0$ and an n-qubit CPTP map $\mathcal{C}$. If there exists $\varepsilon_1, \ldots, \varepsilon_n > 0$, such that $\mathcal{C}$ is a weak ε_i-approximate local identity on the i-th qubit for all $i \in \{1, \ldots, n\}$, then $\mathcal{D}_{\mathrm{ave}}(\mathcal{C}, \mathcal{I}) \leq \frac{3}{2}\sum_{i=1}^{n} \varepsilon_i$.*

Proof of Lemma 5 and 6. Let us define $|\Omega_1\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, and $|\Omega_n\rangle = |\Omega_1\rangle^{\otimes n}$. From Fact 2 characterizing the average-case distance $\mathcal{D}_{\mathrm{ave}}$, we have

$$\mathcal{D}_{\mathrm{ave}}(\mathcal{C}, \mathcal{I}) = \frac{2^n}{2^n + 1}\left(1 - \langle\Omega_n| \left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)|\Omega_n\rangle\right). \tag{80}$$

We can think of the term $\langle\Omega_n| \left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)|\Omega_n\rangle$ as the probability of getting $|\Omega_1\rangle$ on all n parallel two-qubit Bell-basis measurements on the $2n$-qubit state $\left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)$. From standard probability theory, we have the following inequality,

$$1 - \langle\Omega_n| \left(\mathcal{E} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)|\Omega_n\rangle \geq 1 - \mathrm{Tr}\left(\left(|\Omega_1\rangle\langle\Omega_1| \otimes I_{\neq i}^{\otimes 2}\right)\left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)\right), \tag{81}$$

where $|\Omega_1\rangle\langle\Omega_1| \otimes I_{\neq i}^{\otimes 2}$ is a projection onto $|\Omega_1\rangle\langle\Omega_1|$ on the i-th and $(n+i)$-th qubit for any i. Also, from union bound, we have

$$1 - \langle\Omega_n| \left(\mathcal{E} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)|\Omega_n\rangle \leq 1 - \sum_{i=1}^{n}\left(1 - \mathrm{Tr}\left(\left(|\Omega_1\rangle\langle\Omega_1| \otimes I_{\neq i}^{\otimes 2}\right)\left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)\right)\right). \tag{82}$$

By reorganizing using the reduced channel of $\mathcal{C}$ on the i-th qubit, we have

$$\mathrm{Tr}\left(\left(|\Omega_1\rangle\langle\Omega_1| \otimes I_{\neq i}^{\otimes 2}\right)\left(\mathcal{C} \otimes \mathcal{I}\right)\left(|\Omega_n\rangle\langle\Omega_n|\right)\right) = \langle\Omega_1| \left(\mathcal{E}_i^{\mathcal{C}} \otimes \mathcal{I}\right)\left(|\Omega_1\rangle\langle\Omega_1|\right)|\Omega_1\rangle. \tag{83}$$

Therefore, we have

$$\frac{3}{2} \times \frac{2^n}{2^n + 1}\sum_{i=1}^{n} \mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\mathcal{C}}, \mathcal{I}) \geq \mathcal{D}_{\mathrm{ave}}(\mathcal{C}, \mathcal{I}) \geq \frac{3}{2} \times \frac{2^n}{2^n + 1}\mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\mathcal{C}}, \mathcal{I}). \tag{84}$$

By noting that $\frac{3}{2} \geq \frac{3}{2} \times \frac{2^n}{2^n+1}$ and $\frac{3}{2} \times \frac{2^n}{2^n+1} \geq 1$, we obtain Lemma 5 and 6. □

5 Learning shallow quantum circuits from a classical dataset

In this section, we present algorithms for learning shallow quantum circuits that achieve a small diamond distance. All algorithms in this section use a classical dataset obtained from performing randomized measurements on the unknown shallow quantum circuit (defined below) to classically reconstruct the unknown circuit. The learning algorithms only require classical computation.

Definition 8 (Randomized measurement dataset for an unknown unitary)**.** *The learning algorithm accesses an unknown n-qubit unitary U via a randomized measurement dataset of the following form,*

$$\mathcal{T}_U(N) = \left\{|\psi_\ell\rangle = \bigotimes_{i=1}^{n}|\psi_{\ell,i}\rangle, |\phi_\ell\rangle = \bigotimes_{i=1}^{n}|\phi_{\ell,i}\rangle\right\}_{\ell=1}^{N}. \tag{85}$$

A randomized measurement dataset of size N is constructed by obtaining N samples from the unknown unitary U. One sample is obtained from one experiment given as follows.

1. *Sample an input state* $|\psi_\ell\rangle = \bigotimes_{i=1}^n |\psi_{\ell,i}\rangle$*, which is a product state consisting of uniformly random single-qubit stabilizer states in* stab_1*.*

2. *Apply the unknown unitary* U *to* $|\psi_\ell\rangle$*.*

3. *Measure every qubit of* $U|\psi_\ell\rangle$ *under a random Pauli basis. The measurement collapses the state* $U|\psi_\ell\rangle$ *to a state* $|\phi_\ell\rangle = \bigotimes_{i=1}^n |\phi_{\ell,i}\rangle$*, where* $|\phi_{\ell,i}\rangle$ *is a single-qubit stabilizer state* stab_1*.*

Together, N *queries to* U *construct a dataset* $\mathcal{T}_U(N)$ *with* N *samples. The dataset can be represented efficiently on a classical computer with* $\mathcal{O}(Nn)$ *bits.*

An interesting question is whether quantum learning algorithms that have access to the unknown quantum circuit U could be much more efficient. In Section 6, we present a quantum learning algorithm that achieves the optimal scaling in query complexity and computational time for learning geometrically-local shallow quantum circuits over finite gate sets.

5.1 Results

We present the results for learning general and geometrically-local shallow quantum circuits consisting of two-qubit gates over SU(4) and over a finite gate set using a classical dataset.

5.1.1 Learning general shallow quantum circuits

We consider the problem of learning an n-qubit unitary U created by a general shallow quantum circuit C with arbitrary circuit connectivity, i.e., every qubit can be connected to any other qubit by a quantum gate, and an arbitrary number m of ancilla qubits initialized in $|0^m\rangle$ and ended up in $|0^m\rangle$ after C. Formally, we have the following identity for U,

$$U \otimes |0^m\rangle = C(I_n \otimes |0^m\rangle), \tag{86}$$

where I_n is an identity on n qubits.

We have the following theorems for learning the unknown unitary U. We can see that the sample/query complexity is very similar to learning geometric-local circuits. However, the computational complexity becomes higher, and we can only guarantee a polynomial scaling with system size n. The learning algorithm and proof are given in Section 5.3.

Theorem 5 (Learning general shallow quantum circuits)**.** *Given a failure probability* δ*, an approximation error* ε*, and an unknown* n*-qubit unitary* U *generated by a constant-depth circuit over any two-qubit gates in* $\mathrm{SU}(4)$ *with an arbitrary number of ancilla qubits. With a randomized measurement dataset* $\mathcal{T}_U(N)$ *of size*

$$N = \mathcal{O}\left(\frac{n^2 \log(n/\delta)}{\varepsilon^2}\right), \tag{87}$$

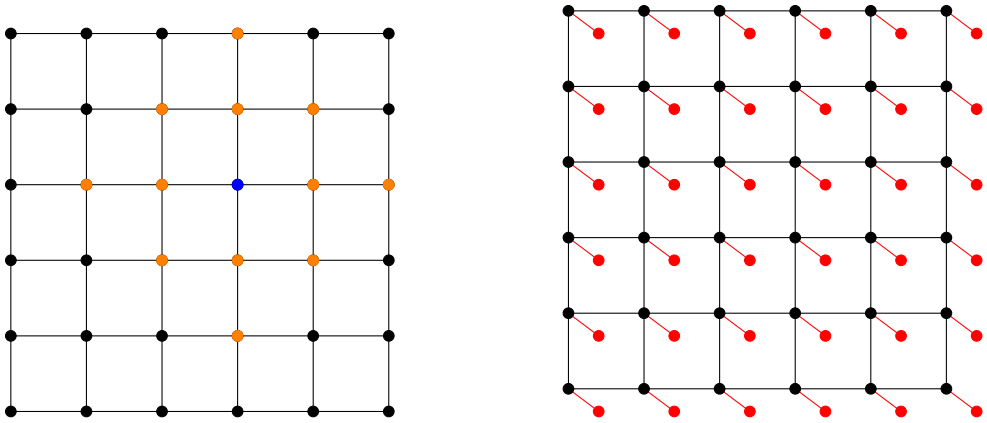
we can learn an n*-qubit quantum channel* $\hat{\mathcal{E}}$ *that can be implemented by a constant-depth quantum circuit over* $2n$ *qubits, such that*

$$\left\|\hat{\mathcal{E}} - \mathcal{U}\right\|_\diamond \leq \varepsilon, \tag{88}$$

with probability at least $1-\delta$*. The classical computational time to learn* $\hat{\mathcal{E}}$ *is* $\mathrm{poly}(n)\log(1/\delta)/\varepsilon^2$*.*

In addition, if each two-qubit gate in the unknown circuit is chosen from a finite gate set of a constant size, then the algorithm learns an exact description $\hat{\mathcal{E}} = \mathcal{U}$ *with probability* $1-\delta$*, using* $N = \mathcal{O}(\log(n/\delta))$ *samples and* $\mathcal{O}(\mathrm{poly}(n)\log(1/\delta))$ *time.*

Remark 1 (Implementation of learned n-qubit channel)**.** *The* n*-qubit channel* $\hat{\mathcal{E}}$ *is the reduced channel* $\mathcal{E}^{\hat{V}}_{\leq n}$ *of the constant-depth* $2n$*-qubit circuit* $\hat{V}$ *on the first* n *qubits.*



(a) Original geometry (b) Learned geometry

Figure 1: Learning geometrically-local shallow quantum circuits. (a) In this example, the geometry is a 2D lattice where each vertex has a degree at most 4. The lightcone of the blue qubit (for depth $d = 2$) is the union of the blue and orange qubits. (b) The learned circuit acts on an extended geometry with $2n$ qubits, where each system qubit (black) is attached to an ancilla qubit (red). Note that each ancilla qubit is connected only with its corresponding system qubit (red edges).

5.1.2 Learning geometrically-local shallow quantum circuits

We consider the problem of learning geometrically-local shallow quantum circuits. Here, we consider a generalized definition of geometric locality, which includes quantum circuits over 1D, 2D, and 3D geometry. The generalization enables more exotic geometry over the qubits and is formally represented by a fixed constant-degree graph. See Fig. 1(a) for an illustration of the definitions.

Definition 9 (Geometric locality)**.** *A geometry over n qubits is defined by a graph $G = (V, E)$ with $n = |V|$ vertices, and each vertex has a degree of at most $\kappa = O(1)$. A geometrically-local two-qubit gate can only act on an edge of G. A geometrically-local quantum circuit is a circuit with only geometrically-local two-qubit quantum gates. A depth-d geometrically-local quantum circuit has d layers, where each layer consists of non-overlapping geometrically-local two-qubit gates.*

Definition 10 (Lightcone in a geometry)**.** *Given a geometry over n qubits represented by a graph $G = (V, E)$ with degree κ and an integer d. The lightcone $L_d(i)$ of a qubit i with depth d is the set of qubits with distance at most d from qubit i in the graph G. We have $|L_d(i)| \leq (\kappa + 1)^d$.*

Definition 11 (Geometrically-local set)**.** *Given a geometry over n qubits represented by a graph $G = (V, E)$. A set S of qubits is geometrically local if all qubits in S are of $\mathcal{O}(1)$ distance in G.*

Under this more general definition of geometry, our proposed algorithm can still learn very efficiently. The following theorem quantifies the efficiency in terms of both the query complexity and the computational complexity. The learning algorithm and proof are given in Section 6.2.

Theorem 6 (Learning geometrically-local shallow quantum circuits)**.** *Given an unknown geometrically local constant-depth n-qubit circuit U over any two-qubit gates in* SU(4)*. With a randomized measurement dataset $\mathcal{T}_U(N)$ of size*

$$N = \mathcal{O}\left(\frac{n^2 \log(n/\delta)}{\varepsilon^2}\right), \tag{89}$$

we can learn an n-qubit quantum channel $\hat{\mathcal{E}}$ that can be implemented by a geometrically local constant-depth quantum circuit over $2n$ qubits, such that

$$\left\|\hat{\mathcal{E}} - \mathcal{U}\right\|_\diamond \leq \varepsilon, \tag{90}$$

with probability at least $1-\delta$. The computational time to learn $\hat{\mathcal{E}}$ is $\mathcal{O}(n^3 \log(n/\delta)/\varepsilon^2)$.

In addition, if each two-qubit gate in the unknown circuit is chosen from a finite gate set of a constant size, then the algorithm learns an exact description $\hat{\mathcal{E}} = \mathcal{U}$ with probability $1-\delta$, using $N = \mathcal{O}(\log(n/\delta))$ samples and $\mathcal{O}(n\log(n/\delta))$ time.

Remark 2 (Implementation of learned n-qubit channel)**.** *The n-qubit channel $\hat{\mathcal{E}}$ is equal to the reduced channel $\mathcal{E}^{\hat{V}}_{\leq n}$ of the geometrically-local constant-depth $2n$-qubit circuit $\hat{V}$ on the first n qubits.*

Next, we look at a result, where we optimize the circuit depth in the learned circuit for implementing $\hat{\mathcal{E}}$. While the depth in the learned circuit can be controlled, the computational complexity becomes substantially worse. The learning algorithm and proof are given in Section 5.5.

Theorem 7 (Learning geometrically-local shallow circuits on k-dimensional lattice with optimized circuit depth)**.** *Given an unknown n-qubit circuit U over any two-qubit gates in* SU(4) *with circuit depth $d = \mathcal{O}(1)$ acting on a k-dimensional lattice with $k = \mathcal{O}(1)$. With a randomized measurement dataset $\mathcal{T}_U(N)$ of size*

$$N = 2^{\mathcal{O}((8kd)^k)} \frac{n^2 \log(n/\delta)}{\varepsilon^2}, \tag{91}$$

we can learn an n-qubit quantum channel $\hat{\mathcal{E}}$ that can be implemented by a quantum circuit over $2n$ qubits on an extended k-dimensional lattice (see Fig. 1(b)), such that

$$\left\|\hat{\mathcal{E}} - \mathcal{U}\right\|_\diamond \leq \varepsilon, \tag{92}$$

with probability at least $1-\delta$.

- *With computational time $\mathcal{O}(n)\cdot N$, the learned circuit has depth at most*

$$(k+1)4^{4(8kd)^k} + 1. \tag{93}$$

- *With computational time $\mathcal{O}(n)\cdot N + (n/\varepsilon)^{\mathcal{O}((8kd)^{k+1})}$, the learned circuit has depth at most*

$$(k+1)(2d+1)+1. \tag{94}$$

In addition, if each two-qubit gate in the unknown circuit is chosen from a finite gate set of a constant size, then the algorithm learns an exact description $\hat{\mathcal{E}} = \mathcal{U}$ with probability $1-\delta$, using $N = \mathcal{O}(\log(n/\delta))$ samples, $\mathcal{O}(n\log(n/\delta))$ time, and a learned circuit of depth $(k+1)(2d+1)+1$.

Remark 3 (The geometry in the doubled system)**.** *In the two theorems given above, we mentioned geometrically-local circuits over $2n$ qubits, while the geometry is defined over n qubits. Given the geometry represented as a graph $G = (V, E)$ over n qubits with $V = \{1, \ldots, n\}$. We extend the graph to $2n$ qubits $G_{\text{ext}} = (V_{\text{ext}}, E_{\text{ext}})$ as follows.*

$$V_{\text{ext}} = \{1, \ldots, n, n+1, \ldots, 2n\}, \quad E_{\text{ext}} = E \cup \{(i, n+i) | 1 \leq i \leq n\}. \tag{95}$$

Each qubit $n+i$ in the added system is connected only to qubit i in the original system; See Fig. 1(b).

5.2 Techniques

We present two sets of closely related techniques for learning an n-qubit unitary U. The first set in Section 5.2.1 uses an idea called local inversion unitary, which follows from the concept of strong approximate local identity given in Section 4. As we have shown earlier, strong local identity checks can be performed by using Heisenberg-evolved single-qubit Pauli observables $U^\dagger P_i U$. The second set in Section 5.2.2 directly uses the Heisenberg-evolved Pauli observables $U^\dagger P_i U$.

5.2.1 Learning using local inversion

We begin by defining the concept of an approximate local inversion unitary.

Definition 12 (Strong ε-approximate local inversion)**.** *Given $n \in \mathbb{N}, \varepsilon \in (0,1)$, $i \in \{1,\ldots,n\}$, and n-qubit unitaries U and V_i. We say V_i is a strong ε-approximate local inversion of U on the i-th qubit if $\mathcal{U}\mathcal{V}_i$ is a strong ε-approximate local identity on the i-th qubit.*

Corollary 2 (Local inversion from Heisenberg-evolved Pauli observables)**.** *Given $n \in \mathbb{N}, \varepsilon \in (0,1)$, $i \in \{1,\ldots,n\}$, and n-qubit unitaries U and V_i. If V_i satisfies*

$$\sum_{P\in\{X,Y,Z\}} \left\| V_i^\dagger U^\dagger P_i U V_i - P_i \right\|_\infty \leq \varepsilon, \tag{96}$$

where P_i acts as $P \in \{X,Y,Z\}$ on the i-th qubit and as identity on the rest of the qubits, then V_i is a strong ε-approximate local inversion of U on the i-th qubit.

Proof. This corollary follows from Lemma 1, which characterizes the strong ε-approximate local identity with Heisenberg evolution of single-qubit Pauli observables. □

Instead of learning the unitary U alone, we consider learning the n local inversion unitaries $V_1,\ldots,V_n$. From the corollary given above, a straightforward way to learn V_i is to first learn the Heisenberg-evolved single-qubit Pauli observable $U^\dagger P_i U$ for all $P = X,Y,Z$, then try to find a unitary V_i that evolves $U^\dagger P_i U$ approximately back to P_i. This could be a much simpler task than learning the entire n-qubit unitary altogether.

While local inversion could potentially make the learning easier, it is *a priori* unclear if learning these local inversions is sufficient to learn U. In the following, we define a formalism for sewing these local inversion unitaries into a $2n$-qubit unitary (instead of n qubits).

Definition 13 (Sewing the local inversions)**.** *Given $n \in \mathbb{N}$ and n-qubit unitaries $V_1,\ldots,V_n$. We define the sewed $2n$-qubit unitary consisting of two sets of n qubits to be the following,*

$$U_{\text{sew}}(V_1,\ldots,V_n) := S\left[\prod_{i=1}^{n}\left(V_i^{(1)}\right) S_i \left(V_i^{(1)}\right)^\dagger\right], \tag{97}$$

where $V_i^{(1)}$ corresponds to applying the n-qubit unitary V_i on the first n qubits, S_i is the swap operator for the i-th qubit between the two sets of n qubits, S is the swap operator for all n qubits.

Remark 4 (Sewing order)**.** *The order for $\left(V_i^{(1)}\right) \mathrm{S}_i \left(V_i^{(1)}\right)^\dagger$ in sewing the local inversions does not matter. We can choose the order to optimize the resulting circuit, e.g., to minimize the circuit depth.*

Lemma 7 (Form of the sewed local inversions)**.** *Given $n \in \mathbb{N}$ and n-qubit unitaries $U, V_1, \ldots, V_n$. Assume V_i is a strong ε_i-approximate local inversion of U on the i-th qubit. Let $U_{\mathrm{sew}} = U_{\mathrm{sew}}(V_1, \ldots, V_n)$.*

$$\mathcal{D}_\diamond(\mathcal{U}_{\mathrm{sew}}, \mathcal{U} \otimes \mathcal{U}^\dagger) = \frac{1}{2}\left\|\mathcal{U}_{\mathrm{sew}} - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond \leq \sum_{i=1}^{n} \varepsilon_i, \tag{98}$$

where the first/second set of n qubits is on the left/right of the tensor product.

Proof. From Theorem 3.55 in [103], we have

$$\left\|\mathcal{U}_{\mathrm{sew}} - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond = \left\|(U^\dagger \otimes U) U_{\mathrm{sew}} \left|\psi\right\rangle\!\left\langle\psi\right| U_{\mathrm{sew}}^\dagger (U \otimes U^\dagger) - \left|\psi\right\rangle\!\left\langle\psi\right|\right\|_1 \tag{99}$$

for some $2n$-qubit state $|\psi\rangle$. We define the following mathematical object,

$$\left|\psi_i\right\rangle\!\left\langle\psi_i\right| := \left[(\mathcal{U}^\dagger \otimes \mathcal{I})\left(\mathcal{S}_1 \ldots \mathcal{S}_i\right)(\mathcal{I} \otimes \mathcal{U})\mathcal{S}\left(\left(\mathcal{V}_{i+1}^{(1)}\right)\mathcal{S}_{i+1}\left(\mathcal{V}_{i+1}^{(1)}\right)^\dagger \ldots \left(\mathcal{V}_n^{(1)}\right)\mathcal{S}_n\left(\mathcal{V}_n^{(1)}\right)^\dagger\right)\right]\left(\left|\psi\right\rangle\!\left\langle\psi\right|\right) \tag{100}$$

for each $i = 0, \ldots, n$. Note that we have the following identities,

$$\left|\psi_0\right\rangle\!\left\langle\psi_0\right| = (U^\dagger \otimes U) U_{\mathrm{sew}} \left|\psi\right\rangle\!\left\langle\psi\right| U_{\mathrm{sew}}^\dagger (U \otimes U^\dagger), \tag{101}$$

$$\left|\psi_n\right\rangle\!\left\langle\psi_n\right| = \left[(\mathcal{U}^\dagger \otimes \mathcal{I})\mathcal{S}(\mathcal{I} \otimes \mathcal{U})\mathcal{S}\right]\left(\left|\psi\right\rangle\!\left\langle\psi\right|\right) = \left|\psi\right\rangle\!\left\langle\psi\right|. \tag{102}$$

By the triangle inequality, we can obtain the following telescoping sum,

$$\left\|\mathcal{U}_{\mathrm{sew}} - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond = \left\|\left|\psi_0\right\rangle\!\left\langle\psi_0\right| - \left|\psi_n\right\rangle\!\left\langle\psi_n\right|\right\|_1 \leq \sum_{i=1}^{n} \left\|\left|\psi_i\right\rangle\!\left\langle\psi_i\right| - \left|\psi_{i-1}\right\rangle\!\left\langle\psi_{i-1}\right|\right\|_1. \tag{103}$$

Each summand can be bounded as follows,

$$\begin{aligned}
\left\|\left|\psi_i\right\rangle\!\left\langle\psi_i\right| - \left|\psi_{i-1}\right\rangle\!\left\langle\psi_{i-1}\right|\right\|_1 &\leq \left\|\mathcal{S}_i(\mathcal{I} \otimes \mathcal{U})\mathcal{S} - (\mathcal{I} \otimes \mathcal{U})\mathcal{S}\left(\mathcal{V}_i \otimes \mathcal{I}\right)\mathcal{S}_i\left(\mathcal{V}_i \otimes \mathcal{I}\right)^\dagger\right\|_\diamond && (104)\\
&= \left\|\mathcal{S}\mathcal{S}_i(\mathcal{U} \otimes \mathcal{I}) - \mathcal{S}(\mathcal{U} \otimes \mathcal{I})\left(\mathcal{V}_i \otimes \mathcal{I}\right)\mathcal{S}_i\left(\mathcal{V}_i \otimes \mathcal{I}\right)^\dagger\right\|_\diamond && (105)\\
&\leq \left\|\mathcal{S}_i(\mathcal{U} \otimes \mathcal{I}) - \left(\left(\mathcal{I}_i \otimes \mathcal{E}_{\neq i}^{\mathcal{U}\mathcal{V}_i}\right) \otimes \mathcal{I}\right)\mathcal{S}_i\left(\mathcal{V}_i \otimes \mathcal{I}\right)^\dagger\right\|_\diamond + \varepsilon_i && (106)\\
&= \left\|\mathcal{S}_i(\mathcal{U} \otimes \mathcal{I}) - \mathcal{S}_i\left(\left(\mathcal{I}_i \otimes \mathcal{E}_{\neq i}^{\mathcal{U}\mathcal{V}_i}\right) \otimes \mathcal{I}\right)\left(\mathcal{V}_i \otimes \mathcal{I}\right)^\dagger\right\|_\diamond + \varepsilon_i && (107)\\
&= \left\|(\mathcal{U}\mathcal{V}_i \otimes \mathcal{I}) - \left(\left(\mathcal{I}_i \otimes \mathcal{E}_{\neq i}^{\mathcal{U}\mathcal{V}_i}\right) \otimes \mathcal{I}\right)\right\|_\diamond + \varepsilon_i \leq 2\varepsilon_i. && (108)
\end{aligned}$$

Together, we obtain the desired statement. □

Remark 5 (A basic identity for $U \otimes U^\dagger$)**.** *A trivial example of an exact local inversion of U on the i-th qubit is $V_i = U^\dagger$. In this case, Lemma 7 yields the following basic identity,*

$$U \otimes U^\dagger = S\left[\prod_{i=1}^{n}\left(U^\dagger \otimes I\right) S_i \left(U \otimes I\right)\right], \tag{109}$$

which can also be shown by canceling all the intermediate $(U \otimes I)\left(U^\dagger \otimes I\right)$.

5.2.2 Learning using Heisenberg-evolved Pauli observables

We have seen earlier that one direct approach to learning local inversion is to first learn the Heisenberg-evolved single-qubit Pauli observables $U^\dagger P_i U$. In the following, we define an alternative formalism that directly sews the Heisenberg-evolved Pauli observables into a $2n$-qubit unitary (instead of n qubits) that approximates $U \otimes U^\dagger$. One can flexibly choose either approach. Typically, learning the Heisenberg-evolved Pauli observables is computationally simpler, but yields higher depth in the learned circuit.

Definition 14 (Approximate Heisenberg-evolved Paui observables)**.** *Given $n \in \mathbb{N}, \varepsilon \in (0,1)$, $i \in \{1,\ldots,n\}$, $P \in \{X,Y,Z\}$, an n-qubit unitary U, and an n-qubit observable $O_{i,P}$. We say $O_{i,P}$ is an ε-approximate Heisenberg-evolved Pauli observable P on qubit i under U if $\left\|O_{i,P} - U^\dagger P_i U\right\|_\infty \leq \varepsilon$.*

Given a set of $3n$ Heisenberg-evolved Pauli observables, we use the following definition to sew them into a $2n$-qubit unitary.

Definition 15 (Sewing the Heisenberg-evolved observables)**.** *Given $n \in \mathbb{N}$ and $3 \times n$ n-qubit observables $O_{i,P}, \forall i = 1,\ldots,n, P \in \{X,Y,Z\}$. Let $\mathrm{Proj}_{\mathrm{U}}(A)$ be the projection of a matrix A to a unitary matrix minimizing the operator norm $\|\cdot\|_\infty$, i.e.,*

$$\mathrm{Proj}_{\mathrm{U}}(A) := \underset{B:unitary}{\arg\min} \|A - B\|_\infty. \tag{110}$$

We define the sewed $2n$-qubit unitary consisting of two sets of n qubits to be the following,

$$U_{\mathrm{sew}}(\{O_{i,P}\}_{i,P}) := S \prod_{i=1}^{n} \left[\mathrm{Proj}_{\mathrm{U}} \left(\frac{1}{2} I \otimes I + \frac{1}{2} \sum_{P \in \{X,Y,Z\}} O_{i,P} \otimes P_i \right) \right], \tag{111}$$

where $V_i^{(1)}$ corresponds to applying the n-qubit unitary V_i on the first n qubits, S_i is the swap operator for the i-th qubit between the two sets of n qubits, S is the swap operator for all n qubits.

Remark 6 (Sewing order)**.** *The order for sewing $\mathrm{Proj}_{\mathrm{U}}\left(\frac{1}{2} I \otimes I + \frac{1}{2}\sum_P O_{i,P} \otimes P_i\right)$ is arbitrary.*

In the above, we have utilized the projection function $\mathrm{Proj}_{\mathrm{U}}$. In the following lemma, we show that this function can be computed efficiently on a classical computer.

Lemma 8 (Projection onto unitary matrices)**.** *Consider the singular value decomposition $A = U\Sigma V^\dagger$, where Σ is diagonal, nonnegative, and U, V is unitary. The projection can be defined as*

$$\mathrm{Proj}_{\mathrm{U}}(A) = UV^\dagger. \tag{112}$$

The computational time is polynomial in the dimension of A.

Proof. Consider any unitary B. We have $\|A - B\|_\infty = \left\|\Sigma - U^\dagger B V\right\|_\infty$. Let W be the unitary $U^\dagger B V$. We can use the definition of $\|M\|_\infty = \sup_v \|Mv\|_2 / \|v\|_2$ to see that

$$\|\Sigma - W\|_\infty \geq \max_i \|\Sigma_{ii}\hat{e}_i - W\hat{e}_i\|_2 \geq \max_i \sqrt{1 + \Sigma_{ii}^2 - 2\Sigma_{ii}\mathrm{Re}[\hat{e}_i^T W \hat{e}_i]} \geq \max_i |1 - \Sigma_{ii}| = \|\Sigma - I\|_\infty, \tag{113}$$

where $\hat{e}_i$ is the unit vector with a nonzero entry on the i-th coordinate. Because $\|\Sigma - I\|_\infty = \left\|A - UV^\dagger\right\|_\infty$, we have obtained $\|A - B\|_\infty \geq \left\|A - UV^\dagger\right\|_\infty$. □

Similar to sewing local inversions, the sewed unitary accurately approximates $U \otimes U^\dagger$.

Lemma 9 (Form of the sewed Heisenberg-evolved observables)**.** *Given $n \in \mathbb{N}$, an n-qubit unitary U, and $3\times n$ n-qubit observables $O_{i,P}, \forall i = 1, \ldots, n, P \in \{X, Y, Z\}$. Assume $O_{i,P}$ is an $\varepsilon_{i,P}$-approximate Heisenberg-evolved Pauli observable P on qubit i under U. Let $U_{\text{sew}} = U_{\text{sew}}(\{O_{i,P}\}_{i,P})$. Then*

$$\mathcal{D}_\diamond(\mathcal{U}_{\text{sew}}, \mathcal{U} \otimes \mathcal{U}^\dagger) = \frac{1}{2}\left\|\mathcal{U}_{\text{sew}} - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond \leq \sum_{i=1}^{n} \sum_{P\in\{X,Y,Z\}} \varepsilon_{i,P}, \tag{114}$$

where the first/second set of n qubits is on the left/right of the tensor product.

Proof. From Eq. (109), we have the following identity,

$$U \otimes U^\dagger = S\left[\prod_{i=1}^{n} (U^\dagger \otimes I) S_i (U \otimes I)\right]. \tag{115}$$

Using the fact that $S_i = \frac{1}{2} I \otimes I + \frac{1}{2}\sum_{P\in\{X,Y,Z\}} P_i \otimes P_i$, we can rewrite the above identity as

$$U \otimes U^\dagger = S \prod_{i=1}^{n}\left[\frac{1}{2} I \otimes I + \frac{1}{2}\sum_{P\in\{X,Y,Z\}} (U^\dagger P_i U) \otimes P_i\right]. \tag{116}$$

Let us denote the following unitaries,

$$V_i := \frac{1}{2} I \otimes I + \frac{1}{2}\sum_{P\in\{X,Y,Z\}} (U^\dagger P_i U) \otimes P_i, \tag{117}$$

$$\widetilde{W}_i := \frac{1}{2} I \otimes I + \frac{1}{2}\sum_{P\in\{X,Y,Z\}} O_{i,P} \otimes P_i \tag{118}$$

$$W_i := \text{Proj}_{\text{U}}\left(\widetilde{W}_i\right). \tag{119}$$

We can upper bound the diamond distance as follows,

$$\begin{aligned}
\left\|\mathcal{U}_{\text{sew}} - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond &= \|\mathcal{V}_n \ldots \mathcal{V}_1 - \mathcal{W}_n \ldots \mathcal{W}_1\|_\diamond && (120)\\
&\leq \sum_{i=1}^{n} \|\mathcal{V}_n \ldots \mathcal{V}_{i+1}\mathcal{W}_i \ldots \mathcal{W}_1 - \mathcal{V}_n \ldots \mathcal{V}_i \mathcal{W}_{i-1} \ldots \mathcal{W}_1\|_\diamond && (121)\\
&\leq \sum_{i=1}^{n} \|\mathcal{V}_n \ldots \mathcal{V}_{i+1}\mathcal{W}_i \ldots \mathcal{W}_1 - \mathcal{V}_n \ldots \mathcal{V}_i \mathcal{W}_{i-1} \ldots \mathcal{W}_1\|_\diamond && (122)\\
&= \sum_{i=1}^{n} \|\mathcal{W}_i - \mathcal{V}_i\|_\diamond \leq 2\sum_{i=1}^{n} \|W_i - V_i\|_\infty. && (123)
\end{aligned}$$

The last inequality uses the fact that $\mathcal{W}_i$ and $\mathcal{V}_i$ are unitary channels. From triangle inequality and the definition of $\text{Proj}_{\text{U}}(\cdot)$, we have the following inequality,

$$\begin{aligned}
\|W_i - V_i\|_\infty &\leq \left\|W_i - \widetilde{W}_i\right\|_\infty + \left\|\widetilde{W}_i - V_i\right\|_\infty \\
&= \min_{V:\text{unitary}} \left\|\widetilde{W}_i - V\right\|_\infty + \left\|\widetilde{W}_i - V_i\right\|_\infty && (124)\\
&\leq 2\left\|\widetilde{W}_i - V_i\right\|_\infty.
\end{aligned}$$

We now use the specific form of $\widetilde{W}_i, V_i$ to upper bound the summand,

$$\|W_i - V_i\|_\infty \leq \sum_{P\in\{X,Y,Z\}} \left\|O_{i,P} - U^\dagger P_i U\right\|_\infty \leq \sum_P \varepsilon_{i,P}. \tag{125}$$

Together with Eq. (123), we can obtain the desired statement. □

Given an n-qubit observable O, we define $\mathrm{supp}(O)$ to be the set of qubits that the observable O acts on. We also define $|O|$ to be the size of $\mathrm{supp}(O)$. We have the following lemma for learning a few-body observable. The learned observable $\hat{O}$ has the property that it only acts on qubits that O acts on, hence $\mathrm{supp}(\hat{O}) \subseteq \mathrm{supp}(O)$.

Lemma 10 (Learning a few-body observable with an unknown support)**.** *Given an error ε, failure probability δ, an unknown n-qubit observable O with $\|O\|_\infty \leq 1$ that acts on an unknown set of k qubits, and a dataset $\mathcal{T}_O(N) = \{|\psi_\ell\rangle = \bigotimes_{i=1}^n |\psi_{\ell,i}\rangle, v_\ell\}_{\ell=1}^N$, where $|\psi_{\ell,i}\rangle$ is sampled uniformly from* stab_1 *and v_ℓ is a random variable with $\mathbb{E}[v_\ell] = \langle\psi_\ell| O |\psi_\ell\rangle$, $|v_\ell| = \mathcal{O}(1)$. Given a dataset size of*

$$N = \frac{2^{\mathcal{O}(k)}\log(n/\delta)}{\varepsilon^2}, \tag{126}$$

with probability at least $1-\delta$, we can learn an observable $\hat{O}$ such that $\left\|\hat{O} - O\right\|_\infty \leq \varepsilon$ and $\mathrm{supp}(\hat{O}) \subseteq \mathrm{supp}(O)$*. The computational complexity is $\mathcal{O}(n^k \log(n/\delta)/\varepsilon^2)$.*

Proof. Consider the observable O under the Pauli basis, $O = \sum_P \alpha_P P$. The α_P coefficients satisfy

$$\alpha_P = 3^{|P|} \mathop{\mathbb{E}}_{|\psi\rangle\sim\mathrm{stab}_1^{\otimes n}} \langle\psi|O|\psi\rangle \, \langle\psi|P|\psi\rangle, \tag{127}$$

which can be learned by replacing the expectation with averaging over the dataset.

We begin by defining the learned observable $\hat{O}$.

$$\hat{\alpha}_P := \frac{3^{|P|}}{N}\sum_{\ell=1}^N v_\ell \langle\psi_\ell| P |\psi_\ell\rangle, \qquad \forall P \in \{I,X,Y,Z\}^{\otimes n} : |P| \leq k, \tag{128}$$

$$\hat{\beta}_P := \begin{cases} \hat{\alpha}_P, & |\hat{\alpha}_P| \geq 0.5\varepsilon/(2\sqrt{2})^k, \\ 0, & |\hat{\alpha}_P| < 0.5\varepsilon/(2\sqrt{2})^k, \end{cases} \tag{129}$$

$$\hat{O} := \sum_{P\in\{I,X,Y,Z\}^{\otimes n}:|P|\leq k} \hat{\beta}_P P. \tag{130}$$

Because O acts on at most k qubits, $\alpha_P = 0$ for $|P| > k$. From Bernstein's inequality, given a dataset size of

$$N = \frac{2^{\mathcal{O}(k)}\log(n/\delta)}{\varepsilon^2}, \tag{131}$$

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

$$|\alpha_P - \hat{\alpha}_P| < 0.5\varepsilon/(2\sqrt{2})^k, \qquad \forall P \in \{I,X,Y,Z\}^{\otimes n} : |P| \leq k. \tag{132}$$

In the following, we assume the above event holds, which happens with probability at least $1-\delta$. We separately prove the following two statements.

$\mathrm{supp}(\hat{O}) \subseteq \mathrm{supp}(O)$: For a Pauli observable P with $\alpha_P = 0$, we have $|\hat{\alpha}_P| < 0.5\varepsilon/(2\sqrt{2})^k$ from Eq. (132). Hence, $\hat{\beta}_P = 0$. As a result, the set of qubits acted by $\hat{O}$ is a subset of $\mathrm{supp}(O)$.

$\left\|\hat{O} - O\right\|_\infty \leq \varepsilon$: From the fact that $\alpha_P = 0$ implies $\hat{\beta}_P = 0$, we have

$$\hat{O} - O = \sum_{P \in \{I,X,YZ\}^{\otimes n} : \mathrm{supp}(P) \subseteq \mathrm{supp}(O)} \left(\hat{\beta}_P - \alpha_P\right) P \tag{133}$$

$$= \sum_{Q \in \{I,X,YZ\}^{\otimes k}} \left(\hat{\beta}_{P(Q)} - \alpha_{P(Q)}\right) P(Q), \tag{134}$$

where $P(Q) := Q \otimes I_{\{1,\ldots,n\}\setminus \mathrm{supp}(O)}$ and $k = |\mathrm{supp}(O)|$. Therefore, we can upper bound the spectral norm by

$$\left\|\hat{O} - O\right\|_\infty \leq \left\| \sum_{Q \in \{I,X,YZ\}^{\otimes k}} \left(\hat{\beta}_{P(Q)} - \alpha_{P(Q)}\right) P(Q) \right\|_\infty = \left\| \sum_{Q \in \{I,X,YZ\}^{\otimes k}} \left(\hat{\beta}_{P(Q)} - \alpha_{P(Q)}\right) Q \right\|_\infty . \tag{135}$$

Recall that $\|A\|_\infty \leq \sqrt{\mathrm{Tr}(A^2)}$ for any Hermitian matrix A, we have

$$\left\|\hat{O} - O\right\|_\infty \leq \sqrt{\sum_{Q \in \{I,X,YZ\}^{\otimes k}} \left(\hat{\beta}_{P(Q)} - \alpha_{P(Q)}\right)^2 \mathrm{Tr}(Q^2)} \leq (2\sqrt{2})^k \max_{|P| \leq k} \left|\hat{\beta}_P - \alpha_P\right|. \tag{136}$$

By the triangle inequality and Eq. (132), we have

$$\left|\hat{\beta}_P - \alpha_P\right| \leq \left|\hat{\beta}_P - \hat{\alpha}_P\right| + |\hat{\alpha}_P - \alpha_P| < \varepsilon/(2\sqrt{2})^k, \qquad \forall |P| \leq k. \tag{137}$$

Therefore, we have obtained the desired inequality $\left\|\hat{O} - O\right\|_\infty \leq \varepsilon$. □

Lemma 11 (Learning a few-body observable with a known support)**.** *Given an error ε, failure probability δ, an unknown n-qubit observable O with $\|O\|_\infty \leq 1$ that acts on an known set S of k qubits, and a dataset $\mathcal{T}_O(N) = \{|\psi_\ell\rangle = \bigotimes_{i=1}^n |\psi_{\ell,i}\rangle, v_\ell\}_{\ell=1}^N$, where $|\psi_{\ell,i}\rangle$ is sampled uniformly from* stab_1 *and v_ℓ is a random variable with $\mathbb{E}[v_\ell] = \langle\psi_\ell| O |\psi_\ell\rangle$, $|v_\ell| = \mathcal{O}(1)$. Given a dataset size of*

$$N = \frac{2^{\mathcal{O}(k)} \log(1/\delta)}{\varepsilon^2}, \tag{138}$$

with probability at least $1-\delta$, we can learn an observable $\hat{O}$ such that $\left\|\hat{O} - O\right\|_\infty \leq \varepsilon$ and $\mathrm{supp}(\hat{O}) \subseteq S$. The computational complexity is $\mathcal{O}(2^{\mathcal{O}(k)} \log(1/\delta)/\varepsilon^2)$.

Proof. We begin by defining the learned observable $\hat{O}$.

$$\hat{\alpha}_P := \frac{3^{|P|}}{N} \sum_{\ell=1}^N v_\ell \langle\psi_\ell| P |\psi_\ell\rangle, \qquad \forall P \in \{I,X,Y,Z\}^{\otimes n} : \mathrm{supp}(P) \subseteq S, \tag{139}$$

$$\hat{O} := \sum_{P \in \{I,X,Y,Z\}^{\otimes n} : \mathrm{supp}(P) \subseteq S} \hat{\alpha}_P P. \tag{140}$$

By definition, we can see that $\mathrm{supp}(\hat{O}) \subseteq S$. Consider the observable O under the Pauli basis, $O = \sum_P \alpha_P P$. Because O acts on the qubits in the set S, $\alpha_P = 0$ for $\mathrm{supp}(P) \not\subseteq S$. From Bernstein's inequality, given a dataset of size

$$N = \frac{2^{\mathcal{O}(k)} \log(1/\delta)}{\varepsilon^2}, \tag{141}$$

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

$$|\alpha_P - \hat{\alpha}_P| < \varepsilon/(2\sqrt{2})^k, \qquad \forall P \in \{I, X, Y, Z\}^{\otimes n} : \mathrm{supp}(P) \subseteq S. \tag{142}$$

In the following, we assume the above event holds, which happens with probability at least $1-\delta$. Using the same derivation as in Eq. (133) to Eq. (136) for the proof of Lemma 10, we have

$$\left\| \hat{O} - O \right\|_\infty \leq (2\sqrt{2})^k \max_{P:\mathrm{supp}(P)\subseteq S} |\hat{\alpha}_P - \alpha_P| < \varepsilon, \tag{143}$$

hence we have arrived at the desired statement. □

Remark 7 (Relation to learning quantum juntas)**.** *The two lemmas given above are related to quantum junta learning [104] but consider a much weaker access model. [104] requires that the unknown observable O be a unitary, and the learning algorithm can access the unitary coherently. In particular, [104] requires inputting half of the maximally entangled state to the unitary. Here, we consider access to O through a simple classical dataset consisting of random product input states and the outcome when measuring the input states with observable O. When the lemmas are used as a subroutine in learning algorithms given in Section 5, we do not have access to O as a unitary, so [104] cannot be used.*

5.3 Learning general shallow circuits (Proof of Theorem 5)

We present the algorithm for learning an unknown n-qubit unitary U generated by an arbitrary constant-depth quantum circuit C with arbitrarily many ancilla qubits. We separate the proof into two-qubit gates over SU(4) and over a finite gate set.

5.3.1 Arbitrary SU(4) gates

The algorithm utilizes a randomized measurement dataset $\mathcal{T}_U(N)$. The key ideas are using Lemma 10 to learn approximate Heisenberg-evolved Pauli observables, using Lemma 13 to sew the Heisenberg-evolved Pauli observables into a constant-depth quantum circuit, and using Lemma 9 to obtain the rigorous performance guarantee.

The following lemma shows how to reuse the randomized measurement dataset $\mathcal{T}_U(N)$ to create the datasets needed to learn approximate Heisenberg-evolved Pauli observables using Lemma 10.

Lemma 12 (Reusing the randomized measurement dataset)**.** *Given an unknown n-qubit unitary U, and a randomized measurement dataset $\mathcal{T}_U(N)$ given in Eq. (85). We can create $3n$ datasets $\mathcal{T}_{U^\dagger P_i U}(N)$, for each Pauli observable $P \in \{X, Y, Z\}$ and each qubit i,*

$$\mathcal{T}_{U^\dagger P_i U}(N) := \left\{ |\psi_\ell\rangle = \bigotimes_{j=1}^n |\psi_{\ell,j}\rangle, v_\ell^{U^\dagger P_i U} \right\}_{\ell=1}^N, \tag{144}$$

where $|\psi_{\ell,i}\rangle$ is sampled uniformly and independently from stab_1 *and $v_\ell^{U^\dagger P_i U}$ is a random variable with $\mathbb{E}[v_\ell^{U^\dagger P_i U}] = \langle\psi_\ell| U^\dagger P_i U |\psi_\ell\rangle$ and $|v_\ell^{U^\dagger P_i U}| = \mathcal{O}(1)$.*

Proof. Recall that from Eq. (85), we have

$$\mathcal{T}_U(N) = \left\{ |\psi_\ell\rangle = \bigotimes_{i=1}^{n} |\psi_{\ell,i}\rangle \,, |\phi_\ell\rangle = \bigotimes_{i=1}^{n} |\phi_{\ell,i}\rangle \right\}_{\ell=1}^{N}. \tag{145}$$

The input states are reused over the $3n$ datasets. For each Pauli observable $P \in \{X, Y, Z\}$ and each qubit i, we define the output value to be

$$v_\ell^{U^\dagger P_i U} := 3 \langle \phi_{\ell,i}| P |\phi_{\ell,i}\rangle \,. \tag{146}$$

We have $|v_\ell^{U^\dagger P_i U}| = |3 \langle \phi_{\ell,i}| P |\phi_{\ell,i}\rangle| \leq 3 = \mathcal{O}(1)$. Now, recall how $|\phi_{\ell,i}\rangle$ is defined. $|\phi_{\ell,i}\rangle$ is the measurement outcome when we measure the i-th qubit of the n-qubit state $U|\psi_\ell\rangle$ in a random Pauli basis: X basis gives $|X,0\rangle := |+\rangle, |X,1\rangle := |-\rangle$; Y basis gives $|Y,0\rangle := |y+\rangle, |Y,1\rangle := |y-\rangle$; Z basis gives $|Z,0\rangle := |0\rangle, |Z,1\rangle := |1\rangle$. Using the fact that

$$0 = \langle Q,b| P |Q,b\rangle \,, \qquad \forall P \neq Q \in \{X,Y,Z\}, b \in \{0,1\}, \tag{147}$$

$$P = \sum_{b\in\{0,1\}} (-1)^b |P,b\rangle\langle P,b| \,, \qquad \forall P \in \{X,Y,Z\}. \tag{148}$$

and that the randomized measurement measures X, Y, Z bases equally likely, we have

$$\mathbb{E}\left[3 \langle \phi_{\ell,i}| P |\phi_{\ell,i}\rangle\right] = \langle \psi_\ell| U^\dagger P_i U |\psi_\ell\rangle \,. \tag{149}$$

This concludes the proof. □

From Lemma 14 and the fact that $\mathrm{supp}\left(U^\dagger P_i U\right) \subseteq A(i) = \bigcup_{P\in\{X,Y,Z\}} \mathrm{supp}\left(U^\dagger P_i U\right)$, we have

$$\left|\mathrm{supp}\left(U^\dagger P_i U\right)\right| \leq |A(i)| = \mathcal{O}(1). \tag{150}$$

This enables us to combine Lemma 12 for constructing $\mathcal{T}_{U^\dagger P_i U}(N), \forall i, P$ from $\mathcal{T}_U(N)$ and Lemma 10 for learning few-body observables with unknown supports (since $A(i)$ is unknown) to show the following. For any constant value $\tilde{\varepsilon} = \mathcal{O}(1)$, given a dataset size of

$$N = \mathcal{O}\left(\frac{n^2 \log(n/\delta)}{\varepsilon^2}\right), \tag{151}$$

we can learn $\hat{O}_{i,P}, \forall i, P$, such that with probability at least $1-\delta$, for all $i \in \{1,\ldots,n\}$ and Pauli observable $P \in \{X,Y,Z\}$, we have

$$\left\|\hat{O}_{i,P} - U^\dagger P_i U\right\|_\infty \leq \frac{\varepsilon}{6n}, \quad \text{and} \quad \mathrm{supp}(\hat{O}_{i,P}) \subseteq \mathrm{supp}(U^\dagger P_i U) \subseteq A(i). \tag{152}$$

The computational time for learning all $\hat{O}_{i,P}$ is $\mathcal{O}(n^{\mathcal{O}(1)} \log(n/\delta)/\varepsilon^2) = \mathrm{poly}(n) \log\left(1/\delta/\varepsilon^2\right)$. From Lemma 14, we can characterize $\mathrm{supp}(\hat{O}_{i,P}) \subseteq \mathrm{supp}(U^\dagger P_i U)$ to apply Lemma 13.

Lemma 13 (Sewing into a constant-depth quantum circuit)**.** *Given $3n$ n-qubit observables $\hat{O}_{i,P}, \forall i \in \{1,\ldots,n\}, P \in \{X,Y,Z\}$, such that for any qubit i, $\left|\bigcup_P \mathrm{supp}\left(\hat{O}_{i,P}\right)\right| = \mathcal{O}(1)$ and there is only a constant number of qubit j with*

$$\left(\bigcup_P \mathrm{supp}\left(\hat{O}_{i,P}\right)\right) \cap \left(\bigcup_P \mathrm{supp}\left(\hat{O}_{j,P}\right)\right) \neq \varnothing. \tag{153}$$

There exists a sewing ordering for $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$ given in Definition 15, such that $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$ can be implemented by a constant-depth quantum circuit. The constant-depth quantum circuit is geometrically-local (see Definition 9) if $\bigcup_P \text{supp}\left(\hat{O}_{i,P}\right), \forall i$ are geometrically-local sets (see Definition 11). The computational time for finding the circuit implementation is $\mathcal{O}(n)$.

Proof. For simplicity of notations, we define $A(i) := \bigcup_P \text{supp}\left(\hat{O}_{i,P}\right)$. We can see that

$$\text{supp}\left(\text{Proj}_{\text{U}}\left(\frac{1}{2}I\otimes I+\frac{1}{2}\sum_{P\in\{X,Y,Z\}}\hat{O}_{i,P}\otimes P_i\right)\right)\subseteq A(i)\cup\{n+i\}, \tag{154}$$

Because $|A(i)| = \left|\bigcup_P \text{supp}\left(\hat{O}_{i,P}\right)\right| = \mathcal{O}(1)$ and Proj_{U} can be implemented in time polynomial in $2^{|A(i)\cup\{n+i\}|} = \mathcal{O}(1)$ as shown in Lemma 8, the following unitary

$$\text{Proj}_{\text{U}}\left(\frac{1}{2}I\otimes I+\frac{1}{2}\sum_{P\in\{X,Y,Z\}}\hat{O}_{i,P}\otimes P_i\right) \tag{155}$$

can be implemented by a constant-depth circuit acting only on qubits in $A(i)\cup\{n+i\}$; see Fact 4 for exact unitary synthesis. Furthermore, if $A(i) = \bigcup_P \text{supp}\left(\hat{O}_{i,P}\right)$ is a geometrically-local set, the constant-depth circuit is geometrically-local; see Corollary 1 for exact unitary synthesis given a connectivity graph. The geometric locality for the $2n$-qubit system is defined in Remark 3.

Consider an n-node graph (equivalently, an n-qubit graph), where each pair (i,j) of nodes (qubits) is connected by an edge if

$$A(i)\cap A(j)\neq\varnothing. \tag{156}$$

The graph only has $\mathcal{O}(n)$ edges and can be constructed as an adjacency list in time $\mathcal{O}(n)$. Because the graph has a constant degree, we can use a $\mathcal{O}(n)$-time greedy graph coloring algorithm to color the n-qubit graph using only a constant number $\chi = \mathcal{O}(1)$ of colors. For each node/qubit i, we consider $c(i)$ to be the color labeled from 1 to χ. The sewing order for the $3n$ observables $\hat{O}_{i,P}$ in Definition 15 are given by the greedy graph coloring, where we order from the smallest color to the largest color. By the definition of graph coloring, for any pair i,j of qubits with the same color, we have

$$A(i)\cap A(j)=\varnothing. \tag{157}$$

Therefore, for any color c', we can find an implementation of the $2n$-qubit unitary

$$\prod_{i:c(i)=c'}\left[\text{Proj}_{\text{U}}\left(\frac{1}{2}I\otimes I+\frac{1}{2}\sum_{P\in\{X,Y,Z\}}O_{i,P}\otimes P_i\right)\right] \tag{158}$$

with a constant-depth (and geometrically-local if $A(i),\forall i$ are geometrically-local) quantum circuit in time $\mathcal{O}(n)$. Since there is only a constant number of colors, the $2n$-qubit unitary $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$ in Eq. (111) with the color-based ordering can be implemented with a constant-depth (and geometrically-local if $A(i),\forall i$ are geometrically-local) quantum circuit in time $\mathcal{O}(n)$. □

Lemma 13 shows that there exists an ordering for sewing the approximate Heisenberg-evolved Pauli observables $\hat{O}_{i,P}$ to create $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$ given in Definition 15, such that $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$

can be implemented by a constant-depth quantum circuit. Given Eq. (152), we can use Lemma 9 on the form of the sewed Heisenberg-evolved Pauli observables to yield

$$\left\|\mathcal{U}_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P}) - \mathcal{U}\otimes\mathcal{U}^\dagger\right\|_\diamond \leq \varepsilon. \tag{159}$$

Finally, define an n-qubit channel $\hat{\mathcal{E}}$ as follows,

$$\hat{\mathcal{E}}(\rho) := \text{Tr}_{>n}\left(\mathcal{U}_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})(\rho\otimes|0^n\rangle\langle 0^n|)\right), \tag{160}$$

which can be implemented as a constant-depth quantum circuit over $2n$ qubits. Because Eq. (152) holds with probability at least $1-\delta$, we have

$$\left\|\hat{\mathcal{E}} - \mathcal{U}\right\|_\diamond \leq \varepsilon \tag{161}$$

with probability at least $1-\delta$. This concludes the proof of the first part of Theorem 5.

5.3.2 Finite gate sets

Let the circuit depth be $d = \mathcal{O}(1)$, the finite gate set be $\mathcal{G}$ with $|\mathcal{G}| = \mathcal{O}(1)$, and the number of ancilla qubits be m. The ancilla qubits are initialized as $|0\rangle$ and end up at $|0\rangle$ after applying C, i.e.,

$$U\otimes|0^m\rangle = C(I_n\otimes|0^m\rangle). \tag{162}$$

The Schrodinger evolution of an n-qubit state ρ under U is

$$U\rho U^\dagger = \text{Tr}_{>n}(C(\rho\otimes|0^m\rangle\langle 0^m|)C^\dagger), \tag{163}$$

where C is a shallow quantum circuit over $n+m$ qubits and $\text{Tr}_{>n}$ traces out the ancilla qubits. The Heisenberg evolution of an n-qubit observable O under U is

$$U^\dagger O U = (I_n\otimes\langle 0^m|)C^\dagger(O\otimes I_m)C(I_n\otimes|0^m\rangle), \tag{164}$$

where I_n is an identity on n qubits and I_m is an identity on m qubits.

The algorithm utilizes a randomized measurement dataset $\mathcal{T}_U(N)$. The key ideas are using Lemma 10 and a brute-force search algorithm over a constant number of choices to find the exact Heisenberg-evolved Pauli observables, using Lemma 13 to sew the Heisenberg-evolved Pauli observables into a constant-depth quantum circuit, and using Lemma 9 to obtain the rigorous guarantee.

Lemma 14 (Characterizing the support)**.** *Given an n-qubit unitary U generated by a constant-depth quantum circuit C with m ancilla qubits. For each qubit $i\in\{1,\ldots,n\}$, let us define a set of qubits*

$$A(i) := \bigcup_{P\in\{X,Y,Z\}} \text{supp}\left(U^\dagger P_i U\right). \tag{165}$$

We have $|A(i)| = \mathcal{O}(1)$ and the number of qubits j such that $A(i)\cap A(j)\neq\varnothing$ is at most a constant.

Proof. From the definition of U, $U\otimes|0^m\rangle = C(I_n\otimes|0^m\rangle)$, we have

$$A(i) \subseteq \bigcup_{P\in\{X,Y,Z\}} \text{supp}\left(C^\dagger P_i C\right). \tag{166}$$

Let $d = \mathcal{O}(1)$ be the depth of the circuit C. We say qubit i is connected to qubit j in the circuit C if there is a sequence of gates in C with strictly decreasing layers, such that each pair of consecutive gates share a qubit and the first gate acts on qubit i and the last gate acts on qubit j. Let $B(i)$ be the set of qubits connected to i. Because each pair of consecutive two-qubit gates share a qubit, the number of possible gate sequences for a fixed i grows at most twice as large at every step. Hence, $|B(i)| \leq 2^d$. Furthermore, for any Pauli operator P, $\text{supp}\left(C^\dagger P_i C\right)$ only contains qubits connected to i, so $A(i) \subseteq B(i)$. Together, $|A(i)| \leq |B(i)| \leq 2^d = \mathcal{O}(1)$. This establishes the first claim.

Now, we show that for any i, the number of j such that $B(i) \cap B(j) \neq \varnothing$ is at most a constant. If $B(i) \cap B(j) \neq \varnothing$, we know that there is a sequence of gates in C with strictly decreasing layers and then strictly increasing layers, such that each pair of consecutive gates share a qubit and the first gate acts on qubit i and the last gate acts on qubit j. Similar to before, The number of possible gate sequences for a fixed i grows at most twice as large at every step. Hence the number of j with $B(i) \cap B(j) \neq \varnothing$ is at most $2^{2d} = \mathcal{O}(1)$. Because $A(i) \subseteq B(i)$, any j with $A(i) \cap A(j) \neq \varnothing$ satisfies $B(i) \cap B(j) \neq \varnothing$. Therefore, the number of qubits j such that $A(i) \cap A(j) \neq \varnothing$ is at most a constant. This establishes the second claim of the lemma. □

From the above lemma and the fact that $\text{supp}\left(U^\dagger P_i U\right) \subseteq A(i)$, we have

$$\left|\text{supp}\left(U^\dagger P_i U\right)\right| \leq |A(i)| = \mathcal{O}(1). \tag{167}$$

This enables us to combine Lemma 12 for constructing $\mathcal{T}_{U^\dagger P_i U}(N), \forall i, P$ from $\mathcal{T}_U(N)$ and Lemma 10 for learning few-body observables with unknown supports (since $A(i)$ is unknown) to show the following. For any constant value $\tilde{\varepsilon} = \mathcal{O}(1)$, given a dataset size of

$$N = \mathcal{O}\left(\log(n/\delta)\right), \tag{168}$$

we can learn $\hat{O}_{i,P}, \forall i, P$, such that with probability at least $1 - \delta$, for all $i \in \{1, \ldots, n\}$ and Pauli observable $P \in \{X, Y, Z\}$, we have

$$\left\|\hat{O}_{i,P} - U^\dagger P_i U\right\|_\infty \leq \tilde{\varepsilon}, \quad \text{and} \quad \text{supp}(\hat{O}_{i,P}) \subseteq \text{supp}(U^\dagger P_i U). \tag{169}$$

The computational time for learning all $\hat{O}_{i,P}$ is $\mathcal{O}(n^{\mathcal{O}(1)} \log(n/\delta)) = \text{poly}(n) \log(1/\delta)$.

Our goal now is to find $U^\dagger P_i U$ exactly using the approximate observable $\hat{O}_{i,P}$ satisfying Eq. (169) by choosing a sufficiently small $\tilde{\varepsilon}$ that is constant in system size n. To do so, we need to consider the backward lightcone of qubit i in circuit C defined below.

Definition 16 (Backward lightcone in a circuit)**.** *We say a gate g in circuit U is in the backward lightcone of qubit i in C if there is a sequence of gates in C with strictly decreasing layers, such that each pair of consecutive gates share a qubit, the first gate acts on qubit i, and the last gate is g.*

The circuit C_i corresponding to the backward lightcone of qubit i in circuit C is the circuit with all gates in the backward lightcone of qubit i in circuit C.

The set S_i of qubits corresponding to the backward lightcone of qubit i in circuit C is the set of all qubits acted by at least one of the gates in the backward lightcone of qubit i in circuit C.

From the definition of C_i, S_i corresponding to the backward lightcones given above, we have

$$\text{supp}(U^\dagger P_i U) \subseteq \text{supp}(C^\dagger P_i C) \subseteq S_i \quad \text{and} \quad U^\dagger P_i U = (I_n \otimes \langle 0^m|) C_i^\dagger P_i C_i (I_n \otimes |0^m\rangle). \tag{170}$$

Note one cannot guarantee $S_i = \text{supp}(U^\dagger P_i U)$. By a counting argument similar to the proof of Lemma 14, we have the following fact.

Fact 5 (Size of backward lightcone)**.** *Given a depth-d circuit C. The circuit C_i corresponding to the backward lightcone of qubit i in C consists of at most 2^{d-1} gates. The set S_i of qubits corresponding to the backward lightcone of qubit i in C contains at most 2^d qubits.*

Recall that the depth of C is $d = \mathcal{O}(1)$, and the gate set is $\mathcal{G}$ with $|\mathcal{G}| = \mathcal{O}(1)$. Because $d = \mathcal{O}(1)$, $|S_i| \leq 2^d = \mathcal{O}(1)$. For any $(n+m)$-qubit constant-depth circuit $\tilde{C}$ over a finite gate set, given a fixed set $\tilde{S}_i$ of qubits corresponding to the backward lightcone of qubit i in $\tilde{C}$, the number of possible circuit $\tilde{C}_i$ corresponding to the backward lightcone of qubit i in circuit $\tilde{C}$ is a constant independent of n, m and $1/\delta$. Hence, there is a constant number of $\tilde{C}_i^\dagger P_i \tilde{C}_i = \tilde{C}^\dagger P_i \tilde{C}$. We denote the possible choices of the n-qubit observable given the set $\tilde{S}_i$ and qubit $i \in \{1, \ldots, n\}$ to be $\mathcal{S}_{\text{obs}}(i, \tilde{S}_i)$,

$$\mathcal{S}_{\text{obs}}(i, \tilde{S}_i) := \Big\{ (I_n \otimes \langle 0^m|) \tilde{C}^\dagger P_i \tilde{C} (I_n \otimes |0^m\rangle) \;\Big|\; \tilde{C} \text{ is a depth-}d \text{ circuit over gate set } \mathcal{G}, \tag{171}$$

$$\text{such that } \tilde{S}_i \text{ is the set of qubits corresponding to the backward lightcone of qubit } i \text{ in } \tilde{C} \Big\} \tag{172}$$

We have $|\mathcal{S}_{\text{obs}}(i, \tilde{S}_i)| = \mathcal{O}(1)$. Furthermore, we can always consider a permutation $\Pi_{i,\tilde{S}_i}$ over the qubits that implements the following permutation mapping,

$$1 \to_{\Pi_{i,\tilde{S}_i}} i, \quad \{1, \ldots, |\tilde{S}_i|\} \to_{\Pi_{i,\tilde{S}_i}} \tilde{S}_i, \tag{173}$$

and $\Pi_{i,\tilde{S}_i}$ acts as identity on the m ancilla qubits. Given a permutation $\Pi_{i,\tilde{S}_i}$ over the qubits (which is itself a unitary), we have

$$\mathcal{S}_{\text{obs}}(i, \tilde{S}_i) = \Big\{ \Pi_{i,\tilde{S}_i} O \Pi_{i,\tilde{S}_i} \;\Big|\; O \in \mathcal{S}_{\text{obs}}(1, \{1, \ldots, |\tilde{S}_i|\}) \Big\}. \tag{174}$$

We note that O acts on n qubits, while $\Pi_{i,\tilde{S}_i}$ acts on $n+m$ qubits; hence, we implicitly extend O to $n+m$ qubits by acting as identity on the m ancilla qubits. The set $\mathcal{S}_{\text{obs}}(1, \{1, \ldots, |\tilde{S}_i|\})$ contains all the possible observables (up to permutation of the qubits) with $|\tilde{S}_i|$ qubits in the backward lightcone of qubit $i \in \{1, \ldots, n\}$ in a depth-d circuit.

Recall from Fact 5 that the set $\tilde{S}_i$ of qubits corresponding to the backward lightcone of qubit i in a depth-d circuit satisfies $1 \leq |\tilde{S}_i| \leq 2^d$. We take the union over all possible values of $|\tilde{S}_i|$ to define

$$\mathcal{S}^*_{\text{obs}} := \bigcup_{k=1}^{2^d} \mathcal{S}_{\text{obs}}(1, \{1, \ldots, k\}). \tag{175}$$

Because $2^d = \mathcal{O}(1)$ and for all $k = \mathcal{O}(1)$, $|\mathcal{S}_{\text{obs}}(1, \{1, \ldots, k\}| = \mathcal{O}(1)$, we have $|\mathcal{S}^*_{\text{obs}}| = \mathcal{O}(1)$. We define the minimum distance between every pair of distinct observables in $\mathcal{S}^*_{\text{obs}}$ as follows,

$$\varepsilon^{\text{dist}} := \min_{O_1 \neq O_2 \in \mathcal{S}^*_{\text{obs}}} \|O_1 - O_2\|_\infty. \tag{176}$$

The minimum distance $\varepsilon^{\text{dist}}$ depends on the depth $d = \mathcal{O}(1)$ and the finite gate set $\mathcal{G}$ with $|\mathcal{G}| = \mathcal{O}(1)$, so ε^* is a constant independent of the system size n and failure probability δ. We also define the minimum distance to an observable with a strictly smaller support.

$$\varepsilon^{\text{supp}} := \min_{O_1 \in \mathcal{S}^*_{\text{obs}}} \min_{\substack{O_2, \text{ such that} \\ \text{supp}(O_2) \subseteq \text{supp}(O_1) \\ \text{supp}(O_2) \neq \text{supp}(O_1)}} \|O_1 - O_2\|_\infty. \tag{177}$$

Because the support of O_2 is strictly contained in the support of O_1, we have $\|O_1 - O_2\|_\infty > 0$. And since $|\mathcal{S}^*_{\text{obs}}| = \mathcal{O}(1)$, we have $\varepsilon^{\text{supp}}$ is a constant independent of n and δ.

Let $\tilde{\varepsilon} = \min(\varepsilon^{\text{dist}}, \varepsilon^{\text{supp}})/3$ in Eq. (169), and define $\hat{S}_i := \{i\} \cup \text{supp}(\hat{O}_{i,P})$. Consider any permutation $\Pi_{i,\hat{S}_i}$ over n qubits that implements the following permutation mapping,

$$1 \to_{\Pi_{i,\hat{S}_i}} i, \quad \{1, \ldots, |\hat{S}_i|\} \to_{\Pi_{i,\hat{S}_i}} \hat{S}_i. \tag{178}$$

We consider the following observable

$$O^*_{i,P} := \Pi_{i,\hat{S}_i} \left(\arg\min_{O \in \mathcal{S}^*_{\text{obs}}} \left\| \Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O \right\|_\infty \right) \Pi_{i,\hat{S}_i}. \tag{179}$$

Because $|\mathcal{S}^*_{\text{obs}}| = \mathcal{O}(1)$ and the dimension of $O \in \mathcal{S}^*_{\text{obs}}$ is a constant, the brute-force minimum over $\mathcal{S}^*_{\text{obs}}$ takes $\mathcal{O}(1)$ time. Because there are $3n$ observables $O^*_{i,P}$, the computational time to find all $3n$ observables $O^*_{i,P}$ is $\mathcal{O}(n)$. The following lemma shows that $O^*_{i,P}$ is exactly equal to the desired Heisenberg-evolved Pauli observable $U^\dagger P_i U$.

Lemma 15 (Exact reconstruction)**.** *Given the definitions above, with probability at least $1 - \delta$, we have $O^*_{i,P} = U^\dagger P_i U$ for all qubits i and Pauli observable P.*

Proof. We condition on the event that Eq. (169) is true, which happens with probability at least $1 - \delta$. Recall that $\text{supp}(\hat{O}_{i,P}) \subseteq \text{supp}(U^\dagger P_i U)$ and $\left\| \hat{O}_{i,P} - U^\dagger P_i U \right\|_\infty \leq \tilde{\varepsilon} \leq \varepsilon^{\text{supp}}/3$. From the definition of $\varepsilon^{\text{supp}}$, we have $\text{supp}(\hat{O}_{i,P}) = \text{supp}(U^\dagger P_i U)$. Hence,

$$\hat{S}_i = \Big(\{i\} \cup \text{supp}(U^\dagger P_i U) \Big) \subseteq S_i, \tag{180}$$

where S_i is the set of qubits corresponding to the backward lightcone of qubit i in circuit C. Consider any permutation $\Pi_{i,\hat{S}_i,S_i}$ over n qubits that is equal to $\Pi_{i,\hat{S}_i}$ for inputs $1, \ldots, |\hat{S}_i|$ and implements the following permutation mapping,

$$\Big\{ |\hat{S}_i| + 1, \ldots, |S_i| \Big\} \to_{\Pi_{i,\hat{S}_i,S_i}} S_i \setminus \hat{S}_i, \tag{181}$$

and $\Pi_{i,\tilde{S}_i,S_i}$ acts as identity on the m ancilla qubits. Because $\text{supp}(U^\dagger P_i U) \subseteq \hat{S}_i$, we have

$$\begin{aligned} \Pi^{-1}_{i,\hat{S}_i} U^\dagger P_i U \Pi^{-1}_{i,\hat{S}_i} &= \Pi^{-1}_{i,\hat{S}_i} (I_n \otimes \langle 0^m|) C^\dagger (P_i \otimes I_m) C (I_n \otimes |0^m\rangle) \Pi^{-1}_{i,\hat{S}_i} \\ &= (I_n \otimes \langle 0^m|) \left(\Pi^{-1}_{i,\hat{S}_i,S_i} C^\dagger \Pi^{-1}_{i,\hat{S}_i,S_i} \right) P_1 \left(\Pi^{-1}_{i,\hat{S}_i,S_i} C \Pi^{-1}_{i,\hat{S}_i,S_i} \right) (I_n \otimes |0^m\rangle). \end{aligned} \tag{182}$$

By the definition of the permutation $\Pi^{-1}_{i,\hat{S}_i,S_i}$, $\{1, \ldots, |S_i|\}$ is the set of qubits corresponding to the backward lightcone of qubit 1 in the circuit $\Pi^{-1}_{i,\hat{S}_i,S_i} C \Pi^{-1}_{i,\hat{S}_i,S_i}$. As a result, we have

$$O^* := (I_n \otimes \langle 0^m|) \left(\Pi^{-1}_{i,\hat{S}_i,S_i} C^\dagger \Pi^{-1}_{i,\hat{S}_i,S_i} \right) P_1 \left(\Pi^{-1}_{i,\hat{S}_i,S_i} C \Pi^{-1}_{i,\hat{S}_i,S_i} \right) (I_n \otimes |0^m\rangle) \tag{183}$$

$$\in \mathcal{S}_{\text{obs}}(1, \{1, \ldots, |S_i|\}) \subseteq \mathcal{S}^*_{\text{obs}}. \tag{184}$$

The last $\subseteq$ follows from the fact that $|S_i| \leq 2^d$ in Fact 5. We can use Eq. (182) and

$$\left\| \hat{O}_{i,P} - U^\dagger P_i U \right\|_\infty \leq \tilde{\varepsilon} \leq \varepsilon^{\text{dist}}/3 \tag{185}$$

to see that

$$\left\| \Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O^* \right\|_\infty \leq \varepsilon^{\text{dist}}/3. \tag{186}$$

For any $O \in \mathcal{S}^*_{\rm obs}$ with $O \neq O^*$, we have $\|O - O^*\|_\infty \geq \varepsilon^{\rm dist}$. By the triangle inequality, we have

$$\left\|\Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O\right\|_\infty \geq \|O - O^*\|_\infty - \left\|\Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O^*\right\|_\infty \geq 2\varepsilon^{\rm dist}/3. \tag{187}$$

Together, we can show that O^* is the unique global minimum,

$$O^* = \arg\min_{O \in \mathcal{S}^*_{\rm obs}} \left\|\Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O\right\|_\infty. \tag{188}$$

Using Eq. (182) again shows that

$$O^*_{i,P} = \Pi_{i,\hat{S}_i} \left(\arg\min_{O \in \mathcal{S}^*_{\rm obs}} \left\|\Pi^{-1}_{i,\hat{S}_i} \hat{O}_{i,P} \Pi^{-1}_{i,\hat{S}_i} - O\right\|_\infty \right) \Pi_{i,\hat{S}_i} = U^\dagger P_i U. \tag{189}$$

This concludes the proof. □

From Lemma 14, we can characterize the support of $O^*_{i,P} = U^\dagger P_i U$ to apply Lemma 13. Lemma 13 shows that there exists an ordering for sewing the Heisenberg-evolved Pauli observables $O^*_{i,P} = U^\dagger P_i U$ to create $U_{\rm sew}(\{O^*_{i,P}\}_{i,P})$ given in Definition 15, such that $U_{\rm sew}(\{O^*_{i,P}\}_{i,P})$ can be implemented by a constant-depth quantum circuit. Under the event that $O^*_{i,P} = U^\dagger P_i U$ (think of $O^*_{i,P}$ as 0-approximate Heisenberg-evolved Pauli observable P on qubit i under U) for all Pauli observable P and qubit i, Lemma 9 shows that

$$U_{\rm sew}(\{O^*_{i,P}\}_{i,P}) = U \otimes U^\dagger. \tag{190}$$

Finally, define an n-qubit channel $\hat{\mathcal{E}}$ as follows,

$$\hat{\mathcal{E}}(\rho) := {\rm Tr}_{>n} \left(\mathcal{U}_{\rm sew}(\{O^*_{i,P}\}_{i,P})(\rho \otimes |0^n\rangle\langle 0^n|) \right), \tag{191}$$

which can be implemented as a constant-depth $2n$ qubits circuit. Using Lemma 15, we have

$$\hat{\mathcal{E}} = \mathcal{U} \tag{192}$$

with probability at least $1 - \delta$. This concludes the proof of Theorem 5.

5.4 Learning geometrically-local shallow circuits (Proof of Theorem 6)

We present the algorithm for learning an unknown geometrically-local shallow quantum circuit U. We separate the proof into two-qubit gates over SU(4) and over a finite gate set.

5.4.1 Arbitrary SU(4) gates

We present the algorithm for learning an unknown geometrically-local shallow quantum circuit U over any two-qubit gate in SU(4). The algorithm uses the randomized measurement dataset $\mathcal{T}_U(N)$. The key ideas are constructing a superset of the support of the Heisenberg-evolved Pauli observables using Lemma 16, finding the Heisenberg-evolved Pauli observables for every qubit using Lemma 11, and sewing the Heisenberg-evolved Pauli observables together using Definition 15 and Lemma 9.

Consider the lightcones $L_d(i)$ for each qubit i with depth d as given in Definition 10. We have the following lemma for characterizing the properties of $L_d(i)$.

Lemma 16 (Properties of lightcones)**.** *Given a geometry over n qubits represented by a graph $G = (V, E)$ with a degree $\kappa = \mathcal{O}(1)$, a depth-d geometrically-local circuit U as given in Definition 9 with $d = \mathcal{O}(1)$, and the lightcones $L_d(i)$ for each qubit i with depth d as given in Definition 10. For each qubit i, we have*

$$\operatorname{supp}\left(U^\dagger P_i U\right) \subseteq L_d(i), \tag{193}$$

for any Pauli operator $P \in \{X, Y, Z\}$. Furthermore, $L_d(i)$ is geometrically local (see Definition 11), $|L_d(i)| = \mathcal{O}(1)$, $L_d(i)$ is known, and the number of qubits j such that $L_d(i) \cap L_d(j) \neq \varnothing$ is at most a constant.

Proof. Because U is of depth d and P_i acts only on qubit i, $U^\dagger P_i U$ only acts only on qubits that are distance d away from qubit i according to the graph G. By the definition of $L_d(i)$, we have $\operatorname{supp}\left(U^\dagger P_i U\right) \subseteq L_d(i)$. Recall that $|L_d(i)| \leq (\kappa+1)^d = \mathcal{O}(1)$. Furthermore, since G is known, $L_d(i)$ is known. Now, consider a qubit j such that $L_d(i) \cap L_d(j) \neq \varnothing$. This condition shows that qubit j must be of distance at most $2d$ from qubit i in the graph G. Hence, the number of such j is bounded above by $(\kappa+1)^{2d} = \mathcal{O}(1)$. This concludes the proof of the lemma. □

Lemma 16 shows that $L_d(i)$ is a geometrically-local set, $|L_d(i)| = \mathcal{O}(1)$, $L_d(i)$ is known, and the number of qubits j such that $L_d(i) \cap L_d(j) \neq \varnothing$ is at most a constant.

Recall that we can use Lemma 12 to constructing $\mathcal{T}_{U^\dagger P_i U}(N), \forall i, P$ from the classical dataset $\mathcal{T}_U(N)$ given in Definition 8. Because $|L_d(i)| = \mathcal{O}(1)$ and $L_d(i)$ is known, from Lemma 11, with a dataset size of

$$N = \mathcal{O}\left(\frac{n^2 \log(3n/\delta)}{\varepsilon^2}\right), \tag{194}$$

we can use $\mathcal{T}_{U^\dagger P_i U}(N), \forall i, P$ constructed from $\mathcal{T}_U(N)$ to learn $\hat{O}_{i,P}, \forall i, P$ such that, with probability at least $1-\delta$, for all $i \in \{1, \ldots, n\}$ and Pauli observable $P \in \{X, Y, Z\}$, we have

$$\left\|\hat{O}_{i,P} - U^\dagger P_i U\right\|_\infty \leq \frac{\varepsilon}{6n} \quad \text{and} \quad \operatorname{supp}(\hat{O}_{i,P}) \subseteq \operatorname{supp}\left(U^\dagger P_i U\right) \subseteq L_d(i). \tag{195}$$

The computational time for learning all $\hat{O}_{i,P}$ is $\mathcal{O}(n^3 \log(n/\delta)/\varepsilon^2)$.

We now utilize Lemm 13 to sew the learned observables into a geometrically-local constant-depth quantum circuit. To use the lemma, we note the following relations from Eq. (195),

$$A(i) := \bigcup_P \operatorname{supp}(\hat{O}_{i,P}) \subseteq \bigcup_P \operatorname{supp}(U^\dagger P_i U) \subseteq L_d(i). \tag{196}$$

Because $L_d(i)$ is a geometrically-local set, $|L_d(i)| = \mathcal{O}(1)$ and the number of qubits j such that $L_d(i) \cap L_d(j) \neq \varnothing$ is at most a constant, we have $A(i)$ is a geometrically-local set, $|A(i)| = \mathcal{O}(1)$ and the number of qubits j such that $A(i) \cap A(j) \neq \varnothing$ is at most a constant. Hence Lemma 13 given above shows that we can find an implementation of $U_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})$ as a geometrically-local constant-depth $2n$-qubit circuit in time $\mathcal{O}(n)$. Given Eq. (195), we can use Lemma 9 on the form of the sewed Heisenberg-evolved Pauli observables to yield

$$\left\|\mathcal{U}_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P}) - \mathcal{U} \otimes \mathcal{U}^\dagger\right\|_\diamond \leq \varepsilon. \tag{197}$$

Finally, define an n-qubit channel $\hat{\mathcal{E}}$ as follows,

$$\hat{\mathcal{E}}(\rho) := \operatorname{Tr}_{>n}\left(\mathcal{U}_{\text{sew}}(\{\hat{O}_{i,P}\}_{i,P})(\rho \otimes |0^n\rangle\langle 0^n|)\right), \tag{198}$$

which can be implemented as a geometrically-local constant-depth quantum circuit over $2n$ qubits. Because Eq. (195) holds with probability at least $1-\delta$, we have

$$\left\|\hat{\mathcal{E}}-\mathcal{U}\right\|_{\diamond}\leq\varepsilon \tag{199}$$

with probability at least $1-\delta$. This concludes the proof of the first part of Theorem 6.

5.4.2 Finite gate sets

We present the algorithm for learning an unknown geometrically-local shallow quantum circuit U over a finite gate set. Let the depth of the unknown shallow quantum circuit be $d=\mathcal{O}(1)$ and the finite gate set be $\mathcal{G}$ with $|\mathcal{G}|=\mathcal{O}(1)$. The algorithm uses the randomized measurement dataset $\mathcal{T}_U(N)$. The algorithm constructs a superset of the support of the Heisenberg-evolved Pauli observables using Lemma 16, finds the Heisenberg-evolved Pauli observables for every qubit exactly using Lemma 11 and the information about the finite gate set $\mathcal{G}$, and sew the Heisenberg-evolved Pauli observables together using Definition 15 and Lemma 9.

Consider the lightcones $L_d(i)$ for each qubit i with depth d as given in Definition 10. Lemma 16 shows that $L_d(i)$ is a geometrically-local set, $|L_d(i)|=\mathcal{O}(1)$, $L_d(i)$ is known, and the number of qubits j such that $L_d(i)\cap L_d(j)\neq\varnothing$ is at most a constant. The algorithm and the proof proceed similarly to the case of having arbitrary two-qubit gates in SU(4). The main difference is in defining the following set $\mathcal{S}_{\text{obs}}(P_i)$ for all $i\in\{1,\ldots,n\}$ and Pauli observable $P\in\{X,Y,Z\}$,

$$\mathcal{S}_{\text{obs}}(P_i):=\left\{U^\dagger P_i U\mid U\text{ is a geometrically-local depth-}d\text{ circuit over the gate set }\mathcal{G}\right\}. \tag{200}$$

Because $|\mathcal{G}|=\mathcal{O}(1)$ and $d=\mathcal{O}(1)$, the set $\mathcal{S}_{\text{obs}}(P_i)$ contains a constant number of observables that only act on qubits in $L_d(i)$. We can define the minimum distance to be

$$\varepsilon_0(P_i):=\min\left\{\ \|O_1-O_2\|_\infty\mid O_1\neq O_2\in\mathcal{S}_{\text{obs}}(P_i)\ \right\}=\Omega(1). \tag{201}$$

We also define $\varepsilon_0=\min_{i,P}\varepsilon_0(P_i)=\Omega(1)$, which is a constant.

Recall that we can use Lemma 12 to constructing $\mathcal{T}_{U^\dagger P_i U}(N),\forall i,P$ from the classical dataset $\mathcal{T}_U(N)$ given in Definition 8. Because $|L_d(i)|=\mathcal{O}(1)$ and $L_d(i)$ is known, from Lemma 11, with a dataset size of

$$N=\mathcal{O}\left(\frac{\log(3n/\delta)}{\varepsilon_0^2}\right)=\mathcal{O}(\log(n/\delta)), \tag{202}$$

we can use $\mathcal{T}_{U^\dagger P_i U}(N),\forall i,P$ constructed from $\mathcal{T}_U(N)$ to learn $\hat{O}_{i,P},\forall i,P$ such that, with probability at least $1-\delta$, for all $i\in\{1,\ldots,n\}$ and Pauli observable $P\in\{X,Y,Z\}$, we have

$$\left\|\hat{O}_{i,P}-U^\dagger P_i U\right\|_\infty\leq\frac{\varepsilon_0}{3}\quad\text{and}\quad\text{supp}(\hat{O}_{i,P})\subseteq\text{supp}\left(U^\dagger P_i U\right)\subseteq L_d(i). \tag{203}$$

The computational time for learning all $\hat{O}_{i,P}$ is $\mathcal{O}(n\log(n/\delta)/\varepsilon_0^2)=\mathcal{O}(n\log(n/\delta))$. Because $U^\dagger P_i U\in\mathcal{S}_{\text{obs}}(P_i)$ only has a constant number of possibilities, we can find

$$O^*_{i,P}:=\operatorname*{arg\,min}_{O\in\mathcal{S}_{\text{obs}}(P_i)}\left\|O-\hat{O}_{i,P}\right\|_\infty \tag{204}$$

in time $\mathcal{O}(n)$. Because the pairwise distance in $\mathcal{S}_{\text{obs}}(P_i)$ is at least ε_0 and $U^\dagger P_i U\in\mathcal{S}_{\text{obs}}(P_i)$,

$$O^*_{i,P}=U^\dagger P_i U,\quad\forall i\in\{1,\ldots,n\},P\in\{X,Y,Z\} \tag{205}$$

with probability at least $1-\delta$.

We now utilize Lemm 13 to sew the learned observables into a geometrically-local constant-depth quantum circuit. To use the lemma, we note the following relations from Eq. (195),

$$A(i) := \bigcup_P \mathrm{supp}(O^*_{i,P}) \subseteq \bigcup_P \mathrm{supp}(U^\dagger P_i U) \subseteq L_d(i). \tag{206}$$

Because $L_d(i)$ is a geometrically-local set, $|L_d(i)| = \mathcal{O}(1)$ and the number of qubits j such that $L_d(i) \cap L_d(j) \neq \varnothing$ is at most a constant, we have $A(i)$ is a geometrically-local set, $|A(i)| = \mathcal{O}(1)$ and the number of qubits j such that $A(i) \cap A(j) \neq \varnothing$ is at most a constant. Hence Lemma 13 given above shows that we can find an implementation of $U_{\mathrm{sew}}(\{O^*_{i,P}\}_{i,P})$ as a geometrically-local constant-depth $2n$-qubit circuit in time $\mathcal{O}(n)$. Given Eq. (205), we can use Lemma 9 on the form of the sewed Heisenberg-evolved Pauli observables to yield

$$\mathcal{U}_{\mathrm{sew}}(\{O^*_{i,P}\}_{i,P}) = \mathcal{U} \otimes \mathcal{U}^\dagger. \tag{207}$$

Finally, define an n-qubit channel $\hat{\mathcal{E}}$ as follows,

$$\hat{\mathcal{E}}(\rho) := \mathrm{Tr}_{>n}\left(\mathcal{U}_{\mathrm{sew}}(\{O^*_{i,P}\}_{i,P})(\rho \otimes |0^n\rangle\langle 0^n|)\right), \tag{208}$$

which can be implemented as a geometrically-local constant-depth quantum circuit over $2n$ qubits. Because Eq. (195) holds with probability at least $1-\delta$, we have

$$\hat{\mathcal{E}} = \mathcal{U} \tag{209}$$

with probability at least $1-\delta$. This concludes the proof of Theorem 6.

5.5 Learning shallow circuits on k-dimensional lattice with optimized circuit depth (Proof of Theorem 7)

Here we develop an approach to optimize the depth of the learned circuit. The main idea is to design a coloring scheme for the k-dimensional lattice with the fewest colors possible, such that gates supported on the same color can be implemented simultaneously.

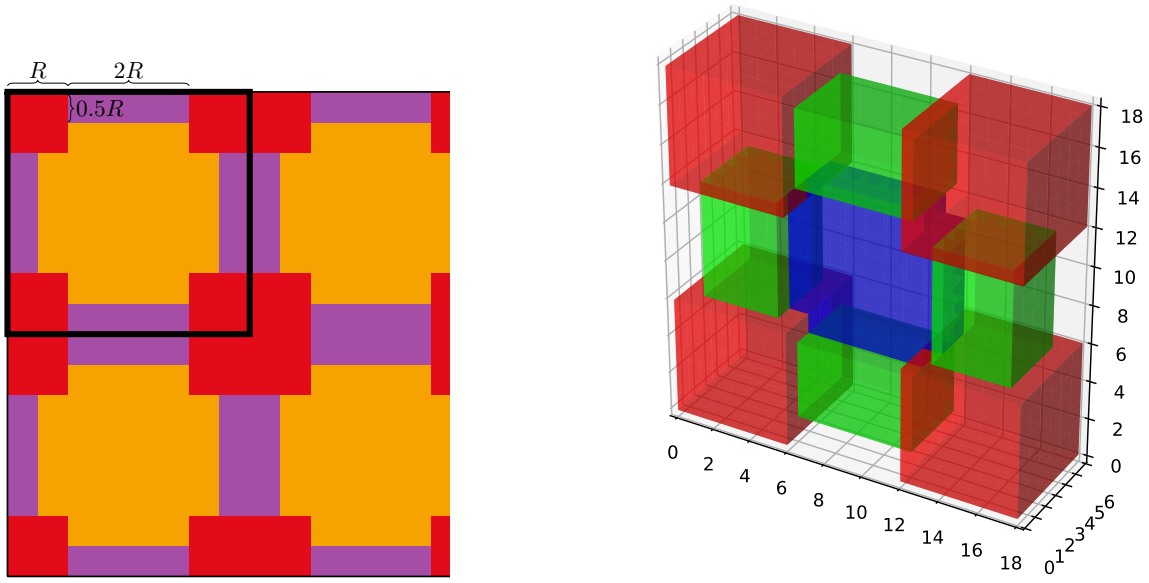
Definition 17 (k+1-coloring of k-dimensional lattice with distance R)**.** *Consider a graph representing a k-dimensional lattice (Fig. 1(a) shows $k=2$). Each vertex is assigned a color, and the entire lattice is divided into many small regions with different colors. A $k+1$-coloring of k-dimensional lattice with distance R satisfies the following properties:*

1. *There are $k+1$ colors in total;*
2. *Each small region has constant size;*
3. *The distance between two regions with the same color is at least R.*

Here we give a construction of the above coloring (see Fig. 2). Similar approaches have been used in e.g. [105], although explicit constructions in 3D or above are not provided. The construction is based on "fattening" different t-cells in the lattice, from small to large t.[1] Consider a k-dimensional cube of length $2kR$ (the volume of the cube is $(2kR)^k$). Then we do the following:

- Fatten each 0-cell (vertices) to length kR, assign color 1.

[1] We thank Jeongwan Haah for teaching this argument at PCMI 2023 Graduate Summer School.



(a) 2D (b) 3D

Figure 2: A coloring of k-dimensional lattice with $k+1$ colors, where different regions of the same color are separated by distance at least R. (a) A coloring of 2-dimensional lattice. (b) A coloring of 3-dimensional lattice (the fourth color is not shown).

- Fatten each 1-cell (edges) to length $(k-1)R$, assign color 2.
- Fatten each 2-cell (faces) to length $(k-2)R$, assign color 3.
- ...
- Fill in the remaining k-cell with color $k+1$.

This is repeated in a translation-invariant way across the entire lattice.

This construction is illustrated in Fig. 2 for $k = 2, 3$. First, consider $k = 2$. A 2-dimensional square of size $4R \times 4R$ is shown in the top left corner (thick black box) of Fig. 2(a). In the first step, we fatten each of the 4 vertices into red squares of size $2R \times 2R$. Only a quarter of each red square remains within the original square. Next, we fatten each of the 4 edges into purple rectangles of size $R \times 2R$. This can be viewed as "growing" the edge until it has thickness R, but the regions that were colored red remain unchanged. Note the fact that the purple edges have a thickness of R, while the red vertices have a thickness of $2R$. This is crucial as it ensures that different purple regions are separated by a distance of at least R. Finally, the remaining regions are colored orange. Note that different orange regions are also separated by a distance of at least R due to the thickness of the purple edges.

The coloring of 3-dimensional lattices is shown in Fig. 2(b). Here we assign colors to a 3-dimensional cube of size $6R \times 6R \times 6R$, and Fig. 2(b) illustrates one of the six faces of that cube, which is the result of fattening the red vertices, green edges, and the blue face (the final coloring of the 3-cell is not shown in the figure). The thickness of the red vertices is larger than the thickness of the green edges, which guarantees that different green edges are separated by distance R. Similarly, the decrease in the thickness of the blue faces relative to the green edges guarantees the separation of different blue faces.

Choose $R = 3d$ in the above coloring scheme, and suppose the system is divided into L small regions $A_1, \ldots, A_L$ ($\sum_i |A_i| = n$). Two regions A_i, A_j that have the same color are separated by distance at least $3d$. Let A_i' be the ancilla system associated with A_i (see Fig. 1), and let S_{A_i} be the SWAP operator across A_i and A_i'. Let $S = \prod_{i=1}^{L} S_{A_i}$ be the global swap between system and ancilla. We are now ready to describe the learning algorithm. We separate the proof into two-qubit gates over SU(4) and over a finite gate set.

5.5.1 Arbitrary SU(4) gates

The learning algorithm proceeds in the same way as in Theorem 6; the only difference is that we need to learn Heisenberg-evolved Pauli operator $U^\dagger P U$ for P supported on each small regions in the coloring scheme instead of on each of the single qubits.

Our goal is to learn to implement the unitary

$$U \otimes U^\dagger = S \left[\prod_{i=1}^{L} (U^\dagger \otimes I) S_{A_i} (U \otimes I) \right]. \tag{210}$$

The algorithm learns each of the operators $W_{A_i} := (U^\dagger \otimes I) S_{A_i} (U \otimes I)$ and then multiply them together, followed by the global swap. The key idea to optimize the circuit depth of the learned circuit is to utilize the coloring scheme in the following sense:

Lemma 17 (Disjointness of supports)**.** *Let A_i, A_j be two regions with the same color. Then W_{A_i} and W_{A_j} have disjoint support.*

Proof. Recall that the operator W_{A_i} is supported on $L(A_i) \cup A_i'$, where $L(A_i)$ is the lightcone of A_i according to Definition 10. Therefore, W_{A_i} does not overlap with W_{A_j} when the lightcones $L(A_i)$ and $L(A_j)$ do not overlap. The coloring scheme has the property that A_i, A_j are separated by distance at least $3d$. Note that the lightcone of a region spreads the region by distance d. This implies that $L(A_i)$ and $L(A_j)$ are still separated by distance at least d and therefore do not overlap. □

Using the above lemma, we can construct the learned circuit by applying the learned operators $\{W_{A_i}\}$ with the same color simultaneously.

Lemma 18. *There is an implementation of $U \otimes U^\dagger$ via applying the operators $\{W_{A_i}\}$ in an appropriate order, such that the total circuit depth is $(k+1)(2d+1)+1$.*

Proof. We would like to implement

$$U \otimes U^\dagger = S \prod_{i=1}^{L} W_{A_i}. \tag{211}$$

Note that the operators $\{W_{A_i}\}$ pairwise commute, and we apply them in the following order: for each color $j \in \{1, 2, \ldots, k+1\}$, apply all operators W_{A_i} that has color j simultaneously. Finally, apply the global swap S.

Note that by definition, $W_{A_i} = (U^\dagger \otimes I) S_{A_i} (U \otimes I)$ can be viewed as a depth-$(2d+1)$ circuit acting on $L(A_i) \cup A_i'$. The total circuit depth is therefore $(k+1)(2d+1)+1$ (the final $+1$ comes from the global swap). □

The learning algorithm has two steps: learning and compiling.

1. (Learning) Learn an approximate classical description $\hat{W}_{A_i}$ for each W_{A_i}, such that $\|\hat{W}_{A_i} - W_{A_i}\|_\infty \leq \varepsilon_1$ for all i with high probability.

2. (Compiling) Compile the learned unitaries $\hat{W}_{A_i}$ from step one into depth-$(2d+1)$ circuits $\hat{W}'_{A_i}$, such that $\|\hat{W}_{A_i} - \hat{W}'_{A_i}\|_\infty \leq 2\varepsilon_1$ for all i.

The diamond distance between the learned circuit and the true circuit is at most $3L\varepsilon_1 \leq 3n\varepsilon_1$.

Step 1: Learning. The goal is to learn an approximation $\hat{O}_{i,P_{A_i}}$ of each operator $U^\dagger P_{A_i} U$, such that the following,

$$\left\|\hat{O}_{i,P_{A_i}} - U^\dagger P_{A_i} U\right\|_\infty \leq \frac{\varepsilon_1}{2^{|A_i|+1}}, \quad \forall i \in \{1,2,\ldots,L\}, \quad P_{A_i} \in \{I,X,Y,Z\}^{|A_i|}, \tag{212}$$

holds with probability at least $1-\delta$.

Using the fact that $S_{A_i} = \frac{1}{2^{|A_i|}} \sum_{P\in\{I,X,Y,Z\}^{|A_i|}} P \otimes P$, we have

$$W_{A_i} = \frac{1}{2^{|A_i|}} \sum_{P\in\{I,X,Y,Z\}^{|A_i|}} U^\dagger P U \otimes P. \tag{213}$$

Meanwhile, let

$$\hat{W}_{A_i} := \mathrm{Proj}_{\mathrm{U}} \left(\frac{1}{2^{|A_i|}} \sum_{P\in\{I,X,Y,Z\}^{|A_i|}} \hat{O}_{i,P_{A_i}} \otimes P_{A_i} \right). \tag{214}$$

From the lattice coloring scheme, we have $|L(A_i)| + |A_i| \leq 2(8kd)^k$. Hence, using Corollary 1 on exact unitary synthesis with geometrically-local circuits, we can implement $\hat{W}_{A_i}$ by a geometrically-local circuit with a circuit depth of

$$4(8kd)^k 4^{2(8kd)^k} \leq 4^{3(8kd)^k+1} \leq 4^{4(8kd)^k}. \tag{215}$$

Conditioned on Eq. (212) succeeds, the approximation error is bounded as follows:

$$\begin{aligned} \left\|\hat{W}_{A_i} - W_{A_i}\right\|_\infty &\leq 2 \left\| \frac{1}{2^{|A_i|}} \sum_{P\in\{I,X,Y,Z\}^{|A_i|}} \left(\hat{O}_{i,P_{A_i}} - U^\dagger P_{A_i} U\right) \otimes P_{A_i} \right\|_\infty \\ &\leq \frac{2}{2^{|A_i|}} \sum_{P\in\{I,X,Y,Z\}^{|A_i|}} \left\|\hat{O}_{i,P_{A_i}} - U^\dagger P_{A_i} U\right\|_\infty \\ &\leq \varepsilon_1. \end{aligned} \tag{216}$$

Here in the first line we use the same argument as in Eq. (124).

It remains to bound the time and query complexity to achieve the learning guarantee in Eq. (212). Given a randomized measurement dataset

$$\mathcal{T}_U(N) = \left\{ |\psi_\ell\rangle = \bigotimes_{i=1}^n |\psi_{\ell,i}\rangle, |\phi_\ell\rangle = \bigotimes_{i=1}^n |\phi_{\ell,i}\rangle \right\}_{\ell=1}^N, \tag{217}$$

for a Pauli operator $P \in \{I,X,Y,Z\}^{|A_i|}$ with weight $w \leq |A_i|$ (the weight of a Pauli operator is the number of non-identity elements), let

$$v_\ell^{U^\dagger P_{A_i} U} := 3^w \langle \phi_{\ell,A_i} | P | \phi_{\ell,A_i} \rangle, \tag{218}$$

where we let $|\phi_{\ell,A_i}\rangle := \otimes_{j\in A_i} |\phi_{\ell,j}\rangle$. The same argument in Lemma 12 shows that

$$\mathbb{E}\left[v_\ell^{U^\dagger P_{A_i} U}\right] = \langle\psi_\ell|U^\dagger P_{A_i} U|\psi_\ell\rangle\,. \tag{219}$$

Let $m := \max_i |L(A_i)| \leq (8kd)^k$ be the maximum support of the operators $U^\dagger P_{A_i} U$. Using Lemma 11, with a dataset size of

$$N = \frac{2^{\mathcal{O}(m)}\log(n/\delta)}{\varepsilon_1^2}, \tag{220}$$

Eq. (212) is achieved with success probability at least $1-\delta$.

Step 2: Compiling. Given a classical description of $\hat{W}_{A_i}$ as unitary acting on $L(A_i)\cup A_i'$, which can be implemented with a circuit depth of at most $4^{4(8kd)^k}$, we would like to find a depth-$(2d+1)$ circuit $\hat{W}'_{A_i}$ that is close to $\hat{W}_{A_i}$. To do this, we construct an ε-net for the circuit lightcone and perform a brute force search.

Definition 18 (ε-net for circuits)**.** *Consider a graph $G=(V,E)$. Let U be some unitary generated by d layers of 2-qubit gates where each gate is chosen from* SU(4) *and acts on an edge in E. An ε-net for circuits is a set of depth-d circuits defined on G, denoted as $\mathcal{N}_\varepsilon(G)$, such that for any choice of U, there exists $V\in\mathcal{N}_\varepsilon(G)$, such that $\|V-U\|_\infty\leq\varepsilon$.*

Lemma 19. *Let $G=(V,E)$ be a graph with $s=|V|$ vertices and maximum degree κ. An ε-net for depth-d circuits defined on G, denoted as $\mathcal{N}_\varepsilon(G)$, can be constructed with size at most $\left(\frac{\kappa sd}{\varepsilon}\right)^{\mathcal{O}(sd)}$ and in time $\left(\frac{\kappa sd}{\varepsilon}\right)^{\mathcal{O}(sd)}$.*

Proof. There are at most $sd/2$ 2-qubit gates in the circuit. We construct the ε-net by first enumerating all possible circuit architectures and then enumerate each 2-qubit gate using a $\frac{2\varepsilon}{sd}$-net for SU(4). In each layer, each qubit can interact with one of the κ neighboring qubits. This implies that the number of possible circuit architectures in one layer is at most κ^s. Therefore, the number of possible circuit architectures with depth d is at most κ^{sd}.

An ε_1-net for SU(4) can be constructed with $\left(\frac{c_0}{\varepsilon_1}\right)^{c_1}$ elements, where c_0, c_1 are absolute constants. Plugging in $\varepsilon_1=\frac{2\varepsilon}{sd}$, the size of $\mathcal{N}_\varepsilon(G)$ is at most

$$\kappa^{sd}\cdot\left(\frac{\mathcal{O}(1)\cdot sd}{\varepsilon}\right)^{\mathcal{O}(1)\cdot sd} = \left(\frac{\kappa sd}{\varepsilon}\right)^{\mathcal{O}(sd)}. \tag{221}$$

This concludes the proof. □

Let $G_{L(A_i)}$ be the subgraph of k-dimensional lattice induced by vertices in $L(A_i)$. The lattice coloring scheme guarantees that the size of $L(A_i)$ is at most $(8kd)^k$. Let $\mathcal{N}_{\varepsilon_2}(G_{L(A_i)})$ be an ε_2-net for depth-d circuits acting on $L(A_i)$, which has size at most

$$\left(\frac{(8kd)^{k+1}}{\varepsilon_2}\right)^{\mathcal{O}(1)\cdot(8kd)^{k+1}}. \tag{222}$$

By definition, there is an element $V\in\mathcal{N}_{\varepsilon_2}(G_{L(A_i)})$ which is a depth-d circuit acting on $L(A_i)$, such that

$$\|(U^\dagger\otimes I)S_{A_i}(U\otimes I)-(V^\dagger\otimes I)S_{A_i}(V\otimes I)\|_\infty\leq 2\varepsilon_2, \tag{223}$$

which implies that

$$\|\hat{W}_{A_i} - (V^\dagger \otimes I) S_{A_i} (V \otimes I)\|_\infty \leq \varepsilon_1 + 2\varepsilon_2. \tag{224}$$

Therefore, enumerating over all elements in $\mathcal{N}_{\varepsilon_2}(G_{L(A_i)})$, we are guaranteed to find one element $\hat{V}$ that satisfies

$$\|\hat{W}_{A_i} - (\hat{V}^\dagger \otimes I) S_{A_i} (\hat{V} \otimes I)\|_\infty \leq \varepsilon_1 + 2\varepsilon_2. \tag{225}$$

Let $\varepsilon_2 = \varepsilon_1/2$ and define $\hat{W}'_{A_i} := (\hat{V}^\dagger \otimes I) S_{A_i} (\hat{V} \otimes I)$, we have $\|\hat{W}_{A_i} - \hat{W}'_{A_i}\|_\infty \leq 2\varepsilon_1$.

Putting everything together. To achieve diamond distance ε between the learned circuit $S \prod_{i=1}^{L} \hat{W}'_{A_i}$ and the true circuit $U \otimes U^\dagger$, it suffices to choose $\varepsilon_1 = \frac{\varepsilon}{3n}$. With probability at least $1 - \delta$, we can learn all operators $\hat{W}_{A_i}$ within sufficient precision, using a dataset size of

$$N = \frac{2^{\mathcal{O}((8kd)^k)} n^2 \log(n/\delta)}{\varepsilon^2}. \tag{226}$$

Next, each $\hat{W}_{A_i}$ is classically compiled into a circuit, and they are combined together according to the order in Lemma 18, such that the learned circuit has total depth $(k+1)(2d+1)+1$. This classical postprocessing procedure takes a total time of

$$\mathcal{O}(nN) + (n/\varepsilon)^{\mathcal{O}(8kd)^{k+1}}, \tag{227}$$

which is polynomial in n and $1/\varepsilon$. If we do not compile $\hat{W}_{A_i}$ to the shorter-depth circuit $\hat{W}'_{A_i}$ and use $\hat{W}_{A_i}$ directly, then the classical postprocessing procedure only requires a computational time of

$$\mathcal{O}(nN), \tag{228}$$

but the learned circuit will have a total depth of $(k+1)4^{4(8kd)^k} + 1$. This concludes the proof of the first part of Theorem 7.

5.5.2 Finite gate sets

The algorithm and the proof closely follow that of arbitrary SU(4) gates. When one considers a finite gate set with a constant size, a key simplification is the following: for any given $i \in \{1, \ldots, L\}$ and $P_{A_i} \in \{I, X, Y, Z\}^{|A_i|}$, $U^\dagger P_{A_i} U$ only takes on a constant number of options. Let $\varepsilon_{i,P_{A_i}} = \Omega(1)$ be the minimum distance in spectral norm between any pair of distinct $U^\dagger P_{A_i} U$.

From the same algorithm and proof in *Step 1: Learning*, we can ensure that

$$\left\| \hat{O}_{i,P_{A_i}} - U^\dagger P_{A_i} U \right\|_\infty \leq \frac{\varepsilon_{i,P_{A_i}}}{3}, \quad \forall i \in \{1, 2, \ldots, L\}, \quad P_{A_i} \in \{I, X, Y, Z\}^{|A_i|}, \tag{229}$$

holds with probability at least $1 - \delta$ using a sample complexity of

$$N = \mathcal{O}\left(\frac{\log(n/\delta)}{\varepsilon^2_{i,P_{A_i}}} \right) = \mathcal{O}\left(\log(n/\delta) \right). \tag{230}$$

From the definition of $\varepsilon_{i,P_{A_i}}$, we can identify $U^\dagger P_{A_i} U$ exactly from $\hat{O}_{i,P_{A_i}}$. This enables us to exactly reconstruct

$$W_{A_i} = \frac{1}{2^{|A_i|}} \sum_{P \in \{I,X,Y,Z\}^{|A_i|}} U^\dagger P U \otimes P = U^\dagger S_{A_i} U. \tag{231}$$

Because U is a quantum circuit of depth $d = \mathcal{O}(1)$ on a constant-dimensional lattice over a finite gate set of a constant size, we can perform a constant-time brute-force search to find a $(2d+1)$-depth circuit implementation for W_{A_i} instead of searching through the ε-net as in *Step 2: Compiling*. The computational time of the compiling step is improved from $(n/\varepsilon)^{\mathcal{O}(8kd)^{k+1})}$ to $\mathcal{O}(n)$. Following the rest of the proof for the case of SU(4) gates, we can learn U exactly with a learned circuit of depth $(k+1)(2d+1)+1$. The sample complexity is given in Eq. (230), and the computational time is dominated by reading the classical dataset, which is of $\mathcal{O}(nN) = \mathcal{O}(n\log(n/\delta))$. This concludes the proof of Theorem 7.

6 Learning shallow quantum circuits from quantum queries

We consider quantum learning algorithms that can access an unknown n-qubit unitary U through coherent quantum queries, which interleave the unitary U with quantum computation.

Definition 19 (Coherent quantum queries)**.** *The learning algorithm is a quantum algorithm with general coherent query access to the unknown unitary U. The quantum learning algorithm can interleave multiple accesses to the unknown unitary U with polynomial-size quantum circuits.*

We show the following result for learning geometrically-local shallow quantum circuits over finite gate sets with asymptotically optimal query complexity and time complexity. We only need to consider proving the matching upper bounds. The matching lower bounds to the query and time complexity are trivial: learning anything about U requires $\Omega(1)$ queries to U; writing down U requires $\Omega(n)$ time.

Theorem 8 (Learning geometrically-local shallow quantum circuits over a finite gate set)**.** *Given an unknown geometrically-local constant-depth n-qubit circuit U over a finite gate set. From*

$$N = \Theta(1) \tag{232}$$

queries to U, we can learn an n-qubit quantum channel $\hat{\mathcal{E}}$ that can be implemented by a geometrically-local constant-depth $2n$-qubit circuit, such that

$$\hat{\mathcal{E}} = \mathcal{U}, \tag{233}$$

with probability 1. The computational time to learn $\hat{\mathcal{E}}$ is $\Theta(n)$.

6.1 Learning local inversion using coherent quantum queries

When there is only a finite choice of possible unitaries, we can find the local inversion perfectly with $\mathcal{O}(1)$ queries, even if there is incoherent noise coming from the environment. This lemma is useful for showing the $\mathcal{O}(1)$ query complexity for learning n-qubit shallow quantum circuits with a finite gate set and a fixed geometric structure. The idea is to store multiple output quantum states in a quantum memory and utilize entangled quantum data processing. The formal statement is given below. We use the subscript on identity I or $\mathcal{I}$ to denote the number of qubits the identity acts on.

Lemma 20 (Perfect local inversion among finite choices)**.** *Consider $k, l, m = \mathcal{O}(1)$, unitaries $U_1, \ldots, U_m$ over k qubits, and unitaries $W_1, \ldots W_m$ over $(k-1)+l$ qubits. Let CPTP maps $\mathcal{E}_x$ from k to $k+l$ qubits be*

$$\mathcal{E}_x(\rho) := (\mathcal{I}_1 \otimes \mathcal{W}_x)(\mathcal{U}_x \otimes \mathcal{I}_l)(\rho \otimes I/2^l), \quad \forall x = 1, \ldots, m. \tag{234}$$

Given an unknown $\mathcal{E}_x$. Using $\mathcal{O}(1)$ queries to $\mathcal{E}_x$, we can find a perfect local inversion V_x of U_x on the first qubit. Furthermore, $V_x = U_i^\dagger$ for some i.

In order to prove the above lemma, we use a perfect local identity check for two choices given in Lemma 21. The proof of Lemma 20 is given after the proof of Lemma 21.

Lemma 21 (Perfect local identity check among two choices)**.** *Consider $k, l \geq 1$, two unitaries U_1, U_2 over k qubits, and two unitaries V_1, V_2 over $k+l-1$ qubits. Given CPTP maps from k qubits to $k+l$ qubits,*

$$\mathcal{E}_x(\rho) := (\mathcal{I}_1 \otimes \mathcal{V}_x)(\mathcal{U}_x \otimes \mathcal{I}_l)(\rho \otimes I/2^l), \quad \forall x = 1, 2. \tag{235}$$

Assume that k, l are constants, U_1 acts as identity on the first qubit $U_1 = I_1 \otimes \tilde{U}_1$, and U_2 is constant far from CPTP maps that act as an identity on the first qubit,

$$c := \min_{\mathcal{E}} \|\mathcal{U}_2 - \mathcal{I}_1 \otimes \mathcal{E}\|_\diamond = \Omega(1). \tag{236}$$

Given an unknown $\mathcal{E}_x$. Using $\mathcal{O}(1)$ queries to $\mathcal{E}_x$, we can perfectly distinguish between $\mathcal{E}_1$ and $\mathcal{E}_2$.

Proof. Let $|\Omega_k\rangle$ be the maximally entangled state over two copies of a k-qubit system. We define the following density matrices over $(k+l)+k$ qubits,

$$\rho_x := (\mathcal{I}_k \otimes \mathcal{E}_x)(|\Omega_k\rangle\langle\Omega_k|), \quad \forall x = 1, 2. \tag{237}$$

The support of a density matrix ρ is defined as

$$\mathrm{supp}(\rho) := \left\{ |\psi\rangle \,\middle|\, \langle\psi| \rho |\psi\rangle > 0 \right\}. \tag{238}$$

From the definition of ρ_x, we have

$$\mathrm{supp}(\rho_x) = \left\{ (I_{k+1} \otimes V_x)(I_k \otimes U_x \otimes I_l)(|\Omega_k\rangle \otimes |\psi\rangle),\ \forall |\psi\rangle \right\}. \tag{239}$$

The maximal fidelity between two density matrices is defined as

$$\tilde{F}(\rho_1, \rho_2) := \max \left(|\langle\phi_1|\phi_2\rangle| \,\middle|\, |\phi_x\rangle \in \mathrm{supp}(\rho_x),\ x = 1, 2 \right). \tag{240}$$

The maximal fidelity behaves similarly to fidelity and is multiplicative under tensor product

$$\tilde{F}(\rho_1 \otimes \sigma_1, \rho_2 \otimes \ \sigma_2) = \tilde{F}(\rho_1, \rho_2)\tilde{F}(\sigma_2, \sigma_2). \tag{241}$$

From the above definition, we see that there exists $|\psi_1\rangle, |\psi_2\rangle$ such that

$$\tilde{F}(\rho_1, \rho_2)^2 = \left| (\langle\Omega_k| \otimes \langle\psi_1|)(I_k \otimes U_2^\dagger \otimes I_l)(I_{k+1} \otimes (V_2^\dagger V_1(\tilde{U}_1 \otimes I_l)))(|\Omega_k\rangle \otimes |\psi_2\rangle) \right|^2. \tag{242}$$

We now consider two states associated with the above,

$$\sigma_1 := (I_{k+1} \otimes (V_2^\dagger V_1(\tilde{U}_1 \otimes I_l)))(|\Omega_k\rangle\langle\Omega_k| \otimes |\psi_2\rangle\langle\psi_2|)(I_{k+1} \otimes ((\tilde{U}_1^\dagger \otimes I_l)V_1^\dagger V_2)) \tag{243}$$

$$\sigma_2 := (I_k \otimes U_2 \otimes I_l)(|\Omega_k\rangle\langle\Omega_k| \otimes |\psi_1\rangle\langle\psi_1|)(I_k \otimes U_2^\dagger \otimes I_l) \tag{244}$$

The Fuchs–van de Graaf inequalities show that $\tilde{F}(\rho_1, \rho_2)^2 = \mathrm{Tr}(\sigma_1\sigma_2) \leq 1 - \frac{1}{4}\|\sigma_1 - \sigma_2\|_1^2$. We now consider a lower bound of the trace norm $\|\sigma_1 - \sigma_2\|_1$ by tracing out the last l qubits,

$$\|\sigma_1 - \sigma_2\|_1 \geq \|(\mathcal{I}_k \otimes \mathcal{I}_1 \otimes \mathcal{E})(|\Omega_k\rangle\langle\Omega_k|) - (\mathcal{I}_k \otimes \mathcal{U}_2)(|\Omega_k\rangle\langle\Omega_k|)\|_1, \tag{245}$$

where $\mathcal{E}$ is a CPTP map that acts on the last $k-1$ qubits. Recall that the 1-norm distance in the Choi states upper bounds the diamond distance in the CPTP maps up to the dimension factor $1/2^k$. From the definition of c in Eq. (236), we have the following inequality,

$$\|\sigma_1 - \sigma_2\|_1 \geq \frac{1}{2^k}\|\mathcal{I}_1 \otimes \mathcal{E} - \mathcal{U}_2\|_\diamond \geq \frac{c}{2^k}. \tag{246}$$

Therefore, we have

$$\tilde{F}(\rho_1, \rho_2) \leq \sqrt{1 - (c/2^{k+2})^2} < 1, \tag{247}$$

which is a key result that will be used later.

We need to consider another pair of states. Consider the Pauli decomposition of U_2 on the first qubit,

$$U_2 = \sum_{P \in \{I,X,Y,Z\}} P \otimes \tilde{U}_{2,P}, \tag{248}$$

where $\tilde{U}_{2,P}$ is a complex matrix of dimension 2^{k-1}. Because U_2 does not act as identity on the first qubit, we have $c' := \sum_{P \neq I} \mathrm{Tr}\left(\tilde{U}_{2,P}^\dagger \tilde{U}_{2,P}\right) > 0$ is a positive constant. Consider the following matrix,

$$M := \sum_{P \in \{X,Y,Z\}} P \otimes \tilde{U}_{2,P}, \tag{249}$$

and define two $2k$-qubit pure states,

$$|\psi_1\rangle := |\Omega_k\rangle\,, \tag{250}$$

$$|\psi_2\rangle := I_k \otimes \left(U_2^\dagger \frac{M}{\sqrt{\mathrm{Tr}(M^\dagger M)/2^k}}\right) |\Omega_k\rangle\,. \tag{251}$$

By the definition of c' and M, we have $\mathrm{Tr}\left(M^\dagger M\right) = 2c' > 0$ and

$$\tilde{F}(|\psi_1\rangle\langle\psi_1|\,, |\psi_2\rangle\langle\psi_2|) = |\,\langle\psi_1|\psi_2\rangle\,|^2 = 2c'/2^k > 0. \tag{252}$$

Furthermore, the overlap between $\mathcal{E}_x(|\psi_x\rangle\langle\psi_x|)$ satisfies

$$\mathrm{Tr}\left(\mathcal{E}_1(|\psi_1\rangle\langle\psi_1|)\mathcal{E}_2(|\psi_2\rangle\langle\psi_2|)\right) = \frac{1}{2c'/2^k} \cdot \frac{1}{2^k} \cdot \frac{1}{2^k}. \tag{253}$$

$$\sum_{P,Q \in \{X,Y,Z\}} \mathrm{Tr}\Big(\mathrm{Tr}_{\leq k}(P \otimes ((\tilde{U}_{2,P}^\dagger \otimes I_l) V_2^\dagger V_1 (\tilde{U}_1 \otimes I_l))) \,\mathrm{Tr}_{\leq k}(Q \otimes ((\tilde{U}_1^\dagger \otimes I_l) V_1^\dagger V_2 (\tilde{U}_{2,Q} \otimes I_l)))\Big) = 0, \tag{254}$$

which implies that there exists a two-outcome projective measurement $\mathcal{M}$ that could perfectly distinguish between the two states $\mathcal{E}_1(|\psi_1\rangle\langle\psi_1|)$ and $\mathcal{E}_2(|\psi_2\rangle\langle\psi_2|)$.

Consider N queries to $\mathcal{E}_x$ to obtain $\rho_x^{\otimes N}$, where the number of queries is

$$N := \max\left(1, \left\lceil \frac{\log\left((2c'/2^k)\right)}{\log\left(\sqrt{1 - (c/2^{k+2})^2}\right)} \right\rceil\right) = \mathcal{O}(1). \tag{255}$$

Using Eq. (241), (247), and (252), we have

$$\tilde{F}(\rho_1^{\otimes N}, \rho_2^{\otimes N}) = \tilde{F}(\rho_1, \rho_2)^N \leq \sqrt{1 - (c/2^{k+2})^2}^{\,N} \leq (2c'/2^k) = \tilde{F}(|\psi_1\rangle\langle\psi_1|\,, |\psi_2\rangle\langle\psi_2|). \tag{256}$$

From Lemma 1 of [98], there exists a CPTP map $\mathcal{T}$ that takes ρ_x to $|\psi_x\rangle\langle\psi_x|$ for $x = 1, 2$. We apply $\mathcal{T}$ to ρ_x. And we evoke one additional query to $\mathcal{E}_x$ to obtain $\mathcal{E}_x(|\psi_x\rangle\langle\psi_x|)$. Finally, we perform the two-outcome projective measurement $\mathcal{M}$ to perfectly distinguish between $\mathcal{E}_1(|\psi_1\rangle\langle\psi_1|)$ and $\mathcal{E}_2(|\psi_2\rangle\langle\psi_2|)$. Together, with $N + 1 = \mathcal{O}(1)$ queries to $\mathcal{E}_x$, we can perfectly distinguish between $\mathcal{E}_1$ and $\mathcal{E}_2$. □

We are now ready to prove Lemma 20. The central idea is a bipartite tournament with a potential local inversion on one side and all possible non-local inversion on the other side.

Proof of Lemma 20. Each query to $\mathcal{E}_x$ allows us to create 1 query to any one of the following CPTP maps,

$$\mathcal{E}_{x,i} = (\mathcal{E}_x \circ \mathcal{U}_i^\dagger),\ \forall i = 1, \ldots, m. \tag{257}$$

The algorithm proceeds by going through all of i one by one. For each i, the algorithm creates two sets,

$$S_i := \left\{ y \in \{1, \ldots, m\} \mid U_y U_i^\dagger \text{ acts as identity on the first qubit} \right\}, \tag{258}$$

$$T_i := \{1, \ldots, m\} \setminus S_i. \tag{259}$$

Note that by definition, $i \in S_i$ and $i \notin T_i$. For each $y \in T_i$, the algorithm uses the algorithm given in the proof of Lemma 21 to test whether $\mathcal{E}_{x,i}$ is equal to $\mathcal{E}_{y,i}$ or $\mathcal{E}_{i,i}$. If $\mathcal{E}_{x,i}$ is indeed equal to one of them, then the algorithm in Lemma 21 is guaranteed to output the one that is equal to $\mathcal{E}_{x,i}$. If not, then the algorithm in Lemma 21 will output $\mathcal{E}_{y,i}$ or $\mathcal{E}_{i,i}$ arbitrarily. After going through all $y \in T_i$, if between $\mathcal{E}_{y,i}$ and $\mathcal{E}_{i,i}$, $\mathcal{E}_{i,i}$ is always chosen for all $y \in T_i$, then the algorithm sets $i^* := i$ and terminates the for-loop over i. The algorithm outputs $U_{i^*}^\dagger$ as the claimed perfect local inversion of U_x on the first qubit.

By construction, the total number of queries to $\mathcal{E}_x$ in the above algorithm is a constant. We now prove that (a) i^* can always be found by the above algorithm and (b) $U_{i^*}^\dagger$ is a perfect local inversion of U_x on the first qubit. The proof is separated into the following two paragraphs addressing each claim.

i^* can always be found. When $i = x$, for each $y \in T_i$, we are testing whether $\mathcal{E}_{x,x}$ is equal to $\mathcal{E}_{y,x}$ or $\mathcal{E}_{x,x}$. Because $U_y U_x^\dagger$ does not act as identity on the first qubit by definition of T_x, Lemma 21 shows that the algorithm will always return $\mathcal{E}_{x,x}$ when deciding between $\mathcal{E}_{y,x}$ and $\mathcal{E}_{x,x}$. Hence when $i = x$, the algorithm will set $i^* := i$ and terminate the for-loop over i. The algorithm could also terminate earlier for some $i < x$ but will always terminate when $i = x$. Therefore, i^*, as defined by the algorithm previously, can always be found.

$U_{i^*}^\dagger$ is a perfect local inversion of U_x on the first qubit. We first show by contradiction that $x \notin T_{i^*}$. Suppose that $x \in T_{i^*}$. For $y = x \in T_{i^*}$, we would be testing whether $\mathcal{E}_{x,i^*}$ is equal to $\mathcal{E}_{x,i^*}$ or $\mathcal{E}_{i^*,i^*}$. Recall that $i^* \notin T_{i^*}$, thus $x \neq i^*$. Lemma 21 thus implies that the algorithm will always return $\mathcal{E}_{x,i^*}$ when deciding between $\mathcal{E}_{x,i^*}$ and $\mathcal{E}_{i^*,i^*}$. As a result, the condition defining i^* is not satisfied, which is a contradiction. Because $S_{i^*} \cup T_{i^*} = \{1, \ldots, m\}$, we have $x \in S_{i^*}$. which means have $U_x U_{i^*}^\dagger$ acts as identity on the first qubit. As a result, $U_{i^*}^\dagger$ is a perfect local inversion of U_x on the first qubit. □

6.2 Learning geometrically-local shallow circuits over a finite gate set (Proof of Theorem 8)

We present the algorithm for learning an unknown geometrically-local shallow quantum circuit U over a finite gate set. Let the geometry over n qubits be represented by a graph $G = (V, E)$ with degree $\kappa = \mathcal{O}(1)$, the depth of U be $d = \mathcal{O}(1)$, and the finite gate set be $\mathcal{G}$ with $|\mathcal{G}| = \mathcal{O}(1)$. This algorithm requires coherent quantum queries to the unknown unitary U. The key ideas are constructing n CPTP maps $\mathcal{E}_i^U, \forall i \in \{1, \ldots, n\}$ from $\mathcal{O}(1)$ queries to U, utilizing Lemma 20 to find

perfect local inversion among finite choices, and using Definition 13 and Lemma 7 to sew the local inversion unitaries together.

We consider the lightcone $L_d(i)$ of the geometry for qubit i under the unknown depth-d geometrically-local circuit U in Definition 10 and the properties of the lightcones given in Lemma 16.

For each qubit i in the n-qubit system, we can always decompose the depth-d geometrically-local quantum circuit U as the following,

$$U = \left(I_i \otimes W^{(i)} \otimes I_{\notin L_{2d}(i)}\right)\left(U^{(i)} \otimes \tilde{W}^{(i)}\right), \tag{260}$$

where $U^{(i)}$ acts on qubits in the set $L_d(i)$, $\tilde{W}^{(i)}$ acts on qubits not in the set $L_d(i)$, $W^{(i)}$ acts on qubits in the set $L_{2d}(i) \setminus \{i\}$, and $I_i, I_{\notin L_{3d}(i)}$ are identity matrices acting on qubit i and qubits not in $L_{3d}(i)$, respectively. Furthermore, $U^{(i)}, W^{(i)}, \tilde{W}^{(i)}$ are all subcircuits (circuits containing a subset of gates) of the unknown depth-d geometrically-local circuits U. We define the CPTP map $\mathcal{E}_i^U$,

$$\mathcal{E}_i^U(\rho) := \mathrm{Tr}_{\notin L_{2d}(i)}\left(U\left(\rho \otimes \frac{I_{\notin L_d(i)}}{2^{n-|L_d(i)|}}\right)U^\dagger\right) \tag{261}$$

$$= \left(\mathcal{I}_i \otimes \mathcal{W}^{(i)}\right)\left(\mathcal{U}^{(i)} \otimes \mathcal{I}_{L_{2d}(i)\setminus L_d(i)}\right)\left(\rho \otimes \frac{I_{L_{2d}(i)\setminus L_d(i)}}{2^{|L_{2d}(i)|-|L_d(i)|}}\right), \tag{262}$$

where ρ is a density matrix for qubits in $L_d(i)$, $I_{\notin L_d(i)}$ is the identity matrix over qubits not in $L_d(i)$, $I_{\notin L_d(i)}/2^{n-|L_d(i)|}$ is the maximally mixed state for qubits not in $L_d(i)$, and $\mathrm{Tr}_{\notin L_{2d}(i)}$ traces out all qubits not in $L_{2d}(i)$. Because $\mathcal{E}_i^U(\rho)$ uses a single query to U, naively, one would expect that to obtain a query to $\mathcal{E}_i^U$ for every qubit i requires n queries to U. The following lemma shows that we can do much more efficiently than what one would naively expect.

Lemma 22 (Queries to every $\mathcal{E}_i^U$ from only $\mathcal{O}(1)$ queries to U)**.** *We can construct a query to every $\mathcal{E}_i^U, 1 \le i \le n$ from only $\mathcal{O}(1)$ queries to the unknown constant-depth geometrically-local circuit U.*

Proof. Let $d = \mathcal{O}(1)$ be the depth of the circuit U. We consider a graph $G^{(3d)}$ over n qubits, where each pair of qubits is connected by an edge if their distance in G is at most $3d$. The degree of $G^{(3d)}$ is at most $(\kappa+1)^{3d} = \mathcal{O}(1)$. The graph only has $\mathcal{O}(n)$ edges and can be constructed as an adjacency list in time $\mathcal{O}(n)$. Let us define a coloring of the graph $G^{(3d)}$. By the standard greedy coloring algorithm, we can find a color $c^{(3d)}(i)$ for each qubit i in graph $G^{(3d)}$, where no adjacent vertices can have the same color, and there are only $\chi^{(3d)}$ distinct colors with

$$\chi^{(3d)} \le (\kappa+1)^{3d} + 1 = \mathcal{O}(1). \tag{263}$$

The greedy coloring algorithm runs in time linear in the number of edges in $G^{(3d)}$, which is linear in the number n of qubits.

For each color $c = 1, \ldots, \chi^{(3d)}$, we consider the set of qubits with color c. We can construct one query to every $\mathcal{E}_i^U$ for qubits i with color $c^{(3d)}(i) = c$ from only one query to U. By the construction of the graph coloring, for two distinct qubits $i \neq j$ with the same color c, $L_{3d}(i) \cap L_{3d}(j) = \varnothing$. We now define the following sets of qubits for the color c,

$$A(c) := \left\{i \in \{1, \ldots, n\} \;\middle|\; c^{(3d)}(i) = c\right\}, \quad B_q(c) := \bigcup_{i: c^{(3d)}(i)=c} L_q(i), \tag{264}$$

for any integer $q \ge 1$. Given the definition of $U^{(i)}, W^{(i)}$ in Eq. (260) for each qubit i. We can further decompose the shallow circuit U as

$$U = \left[\left(I_{A(c)} \otimes \bigotimes_{i: c^{(3d)}(i)=c} W^{(i)}\right) \otimes I_{\notin B_{2d}(c)}\right]\left[\left(\bigotimes_{i: c^{(3d)}(i)=c} U^{(i)}\right) \otimes \tilde{W}^{(c)}\right], \tag{265}$$

where $\tilde{W}^{(c)}$ acts on qubits not in $B_d(c)$. Consider initializing the qubits not in $B_d(c)$ as the maximally mixed state, evolving under U, and tracing out any qubits not in $B_{2d}(c)$. The resulting CPTP map $\mathcal{E}_c^U$ from qubits in $B_d(c)$ to qubits in $B_{2d}(c)$ can be written as

$$\mathcal{E}_c^U(\rho) = \left(\mathcal{I}_{A(c)} \otimes \bigotimes_{i:c^{(3d)}(i)=c} \mathcal{W}^{(i)}\right) \left(\bigotimes_{i:c^{(3d)}(i)=c} \mathcal{U}^{(i)} \otimes \mathcal{I}_{B_{2d}(i)\setminus B_d(i)}\right) \left(\rho \otimes \frac{I_{B_{2d}(c)\setminus B_d(c)}}{2^{|B_{2d}(c)|-|B_d(c)|}}\right), \quad (266)$$

where ρ is a density matrix over qubits in $B_d(c)$. It is not hard to see that

$$\mathcal{E}_c^U = \bigotimes_{i:c^{(3d)}(i)=c} \mathcal{E}_i^U. \quad (267)$$

Because $\mathcal{E}_c^U$ only requires one query to U, we can create $\mathcal{E}_i^U$ for all qubit i with color c from one query to U. Since there is only $\chi^{(3d)} = \mathcal{O}(1)$ colors, we can create a query to every $\mathcal{E}_i^U, 1 \le i \le n$ from only $\mathcal{O}(1)$ queries to the unknown circuit U. □

Because U is over a finite gate set with size $\mathcal{O}(1)$, we have $U^{(i)}$ and $W^{(i)}$ only have a constant number of choices. Furthermore, both $U^{(i)}$ and $W^{(i)}$ act on a constant number of qubits because $|L_d(i)| = \mathcal{O}(1), |L_{2d}(i)| = \mathcal{O}(1)$ for a constant depth d. From Lemma 20, for each qubit i, through $\mathcal{O}(1)$ queries to $\mathcal{E}_i^U$, we can learn a perfect local inversion V_i of $U^{(i)}$ on qubit i with no failure probability. The local inversion unitary V_i is the inverse of one of the possible choices for $U^{(i)}$. Hence, V_i is a geometrically-local depth-d circuit that only acts on qubits in $L_d(i)$. Combining with Lemma 22, from only $\mathcal{O}(1)$ queries to U, we can learn $V^{(i)}, \forall i = 1, \ldots, n$, such that

$$\mathcal{U}^{(i)}\mathcal{V}_i = \mathcal{I}^{(i)} \otimes \mathcal{E}_{\neq i}^{\mathcal{U}^{(i)}\mathcal{V}_i}, \quad (268)$$

where $\mathcal{I}^{(i)}$ is the identity map on qubit i and $\mathcal{E}_{\neq i}^{\mathcal{U}^{(i)}\mathcal{V}_i}$ is the reduced channel of $\mathcal{U}^{(i)}\mathcal{V}_i$ with qubit i removed. The quantum computational time is given by $\mathcal{O}(n)$. We now show that V_i is also the perfect local inversion unitary for U on qubit i. To see this, recall the decomposition in Eq. (260), we have

$$\begin{aligned}
\mathcal{U}\mathcal{V}_i &= \left(\mathcal{I}_i \otimes \mathcal{W}^{(i)} \otimes \mathcal{I}_{\notin L_{2d}(i)}\right) \left(\mathcal{U}^{(i)}\mathcal{V}_i \otimes \tilde{\mathcal{W}}^{(i)}\right) && (269)\\
&= \mathcal{I}^{(i)} \otimes \left(\left(\mathcal{W}^{(i)} \otimes \mathcal{I}_{\notin L_{2d}(i)}\right) \left(\mathcal{E}_{\neq i}^{\mathcal{U}^{(i)}\mathcal{V}_i} \otimes \tilde{\mathcal{W}}^{(i)}\right)\right) && (270)\\
&= \mathcal{I}^{(i)} \otimes \mathcal{E}_{\neq i}^{\mathcal{U}\mathcal{V}_i}. && (271)
\end{aligned}$$

We can now use Definition 13 and Lemma 7 to sew the perfect local inversion unitaries together. This gives the following $2n$-qubit unitary,

$$U_{\text{sew}}(V_1, \ldots, V_n) = S\left[\prod_{i=1}^{n} \left(V_i^{(1)}\right) S_i \left(V_i^{(1)}\right)^\dagger\right] = U \otimes U^\dagger, \quad (272)$$

where $V_i^{(1)}$ is the unitary V_i acting on the first set of n qubits.

We now show that there exists a sewing ordering such that $U_{\text{sew}}(V_1, \ldots, V_n)$ is a constant-depth geometrically-local circuit. Given the geometry over n qubits represented by a graph $G = (V, E)$. Consider a graph $G^{(2d)}$ over n qubits, where each pair (i, j) of qubits are connected by an edge if i, j is of distance at most $2d$ in the geometric graph G. Hence, equivalently, for all (i, j) not connected by an edge in $G^{(2d)}$, we have

$$L_d(i) \cap L_d(j) = \varnothing. \quad (273)$$

The degree of $G^{(2d)}$ is bounded above by $(\kappa+1)^{2d}$. And $G^{(2d)}$ can be constructed as an adjacency list in time $\mathcal{O}(n)$. Because the graph has a constant degree, we can use a $\mathcal{O}(n)$-time greedy graph coloring algorithm to color the n-qubit graph $G^{(2d)}$ using only a constant number of colors. For each node/qubit i, we consider $c(i)$ to be the color. The sewing order for the n local inversion unitaries V_i is given by the greedy graph coloring, where we order from the smallest color to the largest color. By the definition of graph coloring, for any pair i, j of qubits with the same color, we have $L_d(i) \cap L_d(j) = \varnothing$. Furthermore, V_i is a constant-depth geometrically-local circuit that only acts on a constant number of qubits. Therefore, for any color c', we can find an implementation of the $2n$-qubit unitary

$$\prod_{i:c(i)=c'} \left(V_i^{(1)}\right) S_i \left(V_i^{(1)}\right)^\dagger \tag{274}$$

with a constant-depth geometrically-local quantum circuit in time $\mathcal{O}(n)$. Since there is only a constant number of colors, the $2n$-qubit unitary $U_{\text{sew}}(V_1, \ldots, V_n)$ in Eq. (272) with the color-based ordering can be implemented with a constant-depth geometrically-local quantum circuit in time $\mathcal{O}(n)$. Finally, define an n-qubit channel $\hat{\mathcal{E}}$ as follows,

$$\hat{\mathcal{E}}(\rho) := \text{Tr}_{>n}\left(\mathcal{U}_{\text{sew}}(V_1, \ldots, V_n)(\rho \otimes |0^n\rangle\langle 0^n|)\right), \tag{275}$$

which can be implemented as a geometrically-local constant-depth quantum circuit over $2n$ qubits. Because $U_{\text{sew}}(V_1, \ldots, V_n) = U \otimes U^\dagger$ from Eq. (272), we have

$$\mathcal{E} = \mathcal{U} \tag{276}$$

with probability one. This concludes the proof of Theorem 8.

7 Hardness for learning log-depth quantum circuits

We have seen from the previous appendices that learning general constant-depth quantum circuits can be done efficiently. A natural follow-up question is whether one could efficiently learn log-depth quantum circuits. In the following, we show that learning log-depth quantum circuits to a constant diamond distance is exponentially hard, even when we allow coherent quantum queries to U. Hence, the problem of learning quantum circuits transitions from being polynomially easy to exponentially hard when we go from $\mathcal{O}(1)$-depth to $\mathcal{O}(\log n)$-depth.

Proposition 3 (Hardness for learning log-depth circuits)**.** *Consider an unknown n-qubit unitary U generated by a $\mathcal{O}(\log n)$-depth circuit over arbitrary two-qubit gates with n ancilla qubits. We have*

- *Learning U to $1/3$ diamond distance with high probability requires $\exp(\Omega(n))$ queries.*
- *Distinguishing whether U equals to the identity I or is $1/3$-far from the identity I in diamond distance with high probability requires $\exp(\Omega(n))$ queries.*

Proof. Without loss of generality, we consider n to be 2^k for an integer k. Consider the unknown unitary U to be I or one of $U_x, \forall x \in \{0,1\}^n$. The unitary U_x is defined to be

$$U_x |y\rangle = \begin{cases} 1, & x = y, \\ -1, & x \neq y, \end{cases} \tag{277}$$

for any $y \in \{0,1\}^n$. The n-qubit unitary U_x can be constructed as follows,

$$U_x = \left(\prod_{\substack{1 \le i \le n \\ x_i = 0}} X_i \right) C^n Z \left(\prod_{\substack{1 \le i \le n \\ x_i = 0}} X_i \right), \tag{278}$$

where X_i is the X gate on the i-th qubit, and $C^n Z$ is a controlled-Z gate controlled on all qubits. The circuit $\prod_{i:\, x_i = 0} X_i$ can be implemented in one layer. We can implement $C^n Z$ using n ancilla qubits in depth $\mathcal{O}(\log n)$. To see this, we first construct a $(2^k + 2^k - 1)$-qubit unitary V recursively as follows:

1. Set the $n = 2^k$ qubits to be the first set of control qubits. Set $j \leftarrow k$.
2. Consider the 2^j control qubits as 2^{j-1} pairs of two control qubits. Include 2^{j-1} new ancilla qubits initialized at $|0\rangle^n$.
3. For each pair of control qubits, implement a CCX gate on each newly added ancilla qubit controlled on the two control qubits.
4. Set the new 2^{j-1} ancilla qubits as the set of control qubits. Set $j \leftarrow j - 1$.
5. If $j > 0$, repeat Step 2.

We can compile the CCX gate acting on three qubits to be a sequence with a constant number of two-qubit gates. The depth of V is $\mathcal{O}(\log n)$. The unitary V computes whether all n qubits are one and stores the result in the $2n - 1$ qubit. We can implement the n-qubit unitary $C^n Z$ using a $2n$-qubit $\mathcal{O}(\log n)$-depth circuit with n ancilla qubits,

$$C^n Z \otimes |0^n\rangle = (V \otimes I)^\dagger \, X_{2n} \, \mathrm{CZ}_{2n-1,2n} \, X_{2n} \, (V \otimes I) \, (I_n \otimes |0^n\rangle), \tag{279}$$

where X_{2n} is the NOT gate on the one ancilla qubit not acted by V, I is a single-qubit identity, I_n is an n-qubit identity, and $\mathrm{CZ}_{2n-1,2n}$ is controlled on the last ancilla qubit added in the recursive construction of V and acts on the one ancilla qubit not acted by V.

If one could learn U up to $1/3$ error in the diamond distance with high probability or if one could distinguish whether U equals to the identity I or is $1/3$-far from the identity I in the diamond distance with high probability, then one could successfully distinguish between the identity map I and the unitary U_x. Distinguishing I or one of $U_x, \forall x \in \{0,1\}^n$ is the well-known Grover search problem. Hence, from the well-known Grover lower bound [106], we have the number of queries must be at least $\Omega(2^{n/2}) = \exp(\Omega(n))$. This concludes the proof. □

8 Learning quantum states generated by shallow circuits in 2D

Given copies of an unknown quantum state $|\psi\rangle = U|0^n\rangle$, with the promise that U is a depth-d circuit acting on a 2-dimensional lattice. In this section, we present an algorithm to learn a description of a shallow circuit that prepares $|\psi\rangle$ up to a desired precision. The algorithm can be viewed as first collecting a sufficiently large randomized measurement dataset [84, 88] from the unknown state and then classically reconstructing the circuit based on the dataset.

Definition 20 (Randomized measurement dataset for an unknown state)**.** *The learning algorithm accesses the unknown state via a randomized measurement dataset of the following form,*

$$\mathcal{T}_{|\psi\rangle}(N) = \left\{ |\phi_\ell\rangle = \bigotimes_{i=1}^{n} |\phi_{\ell,i}\rangle \right\}_{\ell=1}^{N}. \tag{280}$$

A randomized measurement dataset of size N is constructed by obtaining N samples from the unknown state $|\psi\rangle$. One sample is obtained from one experiment given as follows: measure every qubit of $|\psi\rangle$ under a random Pauli basis. The measurement collapses the state $|\psi\rangle$ to a state $|\phi_\ell\rangle = \bigotimes_{i=1}^{n} |\phi_{\ell,i}\rangle$, where $|\phi_{\ell,i}\rangle$ is a single-qubit stabilizer state in stab_1.

Together, N copies of $|\psi\rangle$ construct a dataset $\mathcal{T}_{|\psi\rangle}(N)$ with N samples. The dataset can be represented efficiently on a classical computer with $\mathcal{O}(Nn)$ bits.

Theorem 9 (Learning quantum states generated by shallow circuits in 2D)**.** *Given copies of an unknown state $|\psi\rangle$, with the promise that $|\psi\rangle = U|0^n\rangle$ for an unknown n-qubit circuit U with circuit depth d acting on a 2-dimensional lattice, then the following holds.*

1. *Suppose each two-qubit gate in U is chosen from* $\mathrm{SU}(4)$. *With a randomized measurement dataset $\mathcal{T}_{|\psi\rangle}(N)$ of size*

$$N = \frac{2^{\mathcal{O}(d^2)} n^{50}}{\varepsilon^{64}} \log \frac{n}{\delta}, \tag{281}$$

we can learn a quantum circuit V with depth $3d$ acting on $n+m$ qubits on an extended 2-dimensional lattice, such that

$$\frac{1}{2} \left\| \mathrm{Tr}_B \left(V |0^n\rangle\langle 0^n|_A \otimes |0^m\rangle\langle 0^m|_B V^\dagger \right) - |\psi\rangle\langle\psi| \right\|_1 \leq \varepsilon, \tag{282}$$

with probability at least $1-\delta$. The computational time to learn V is $\left(\frac{nd^3}{\varepsilon}\right)^{\mathcal{O}(d^3)}$. The number of ancilla qubits can be chosen as $m = tn$ for an arbitrarily small constant $t > 0$.

2. *In addition, if each two-qubit gate in U is chosen from a finite gateset of constant size and $d = \mathcal{O}(1)$, then there is an algorithm that learns an exact preparation circuit V with depth $3d$ acting on $n+m$ qubits, such that $V|0^n\rangle_A |0^m\rangle_B = |\psi\rangle_A \otimes |\mathrm{junk}\rangle_B$ with probability $1-\delta$, with sample complexity $N = \mathcal{O}(\log(n/\delta))$ and time complexity $\mathcal{O}(n \log(n/\delta))$. The number of ancilla qubits can be chosen as $m = tn$ for an arbitrarily small constant $t > 0$.*

3. *In addition, if each two-qubit gate in U is chosen from a finite gateset of constant size and $d = \mathcal{O}(1)$, then there is an algorithm that learns a circuit V with depth $2^{c \cdot d^2}$ (for some universal constant c) acting on n qubits (without using any ancilla), such that $\left|\langle 0^n| V^\dagger |\psi\rangle\right|^2 \geq 1-\varepsilon$ with probability $1-\delta$, with query complexity $N = \mathcal{O}(\log(n/\delta))$ and time complexity $(n/\varepsilon)^{\mathcal{O}(1)}$.*

Remark 8. *The first claim in Theorem 9 holds for any gateset and any circuit depth d (which may not be a constant), while the second and third claims are specialized to the simpler setting of finite gateset and constant depth.*

In particular, the first claim implies that when $d = \mathrm{polylog}(n)$, the state $|\psi\rangle$ can be learned within ε trace distance with sample complexity $N = \frac{2^{\mathrm{polylog}(n)}}{\varepsilon^{\mathcal{O}(1)}} \log \frac{n}{\delta}$, in time $(n/\varepsilon)^{\mathrm{polylog}(n)}$.

We prove Theorem 9 in the remainder of this section. Next we give a detailed presentation of the argument outlined in Section 2.2.1 and 2.2.2. We start by assuming a finite gate set, and address general SU(4) gates in Section 8.4.

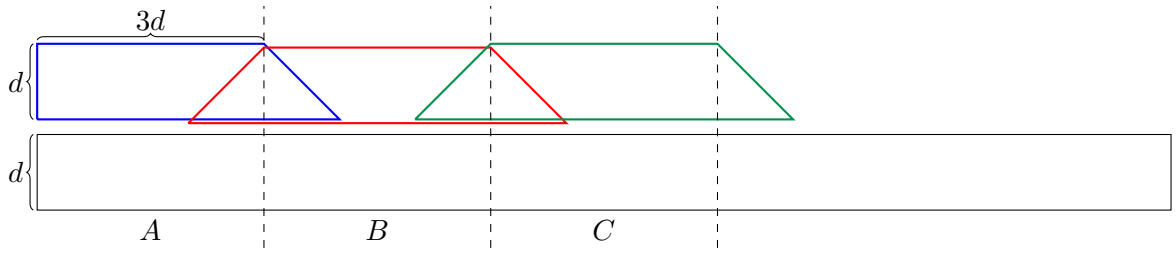


Figure 3: Efficient learning of quantum states generated by a shallow circuit in 1D. For each local region $A, B, C, \ldots$ we find a list of local inversion circuits, and merge them together by solving a constraint satisfaction problem.

8.1 Learning 1D states by solving a constraint satisfaction problem

We start by assuming U is a depth-d circuit acting on a 1D lattice, for some constant $d = \mathcal{O}(1)$. The learning problem is equivalent to finding a low-depth circuit V such that $V|\psi\rangle = |0^n\rangle$. Consider Fig. 3 where A, B, C are contiguous regions of size $3d$. Suppose we want to locally invert the qubits in region A back to $|0\rangle_A$. We can do so by undoing the gates within the lightcone of A, i.e. apply a depth-d circuit of the blue shape (that acts on $4d$ qubits) on top of $|\psi\rangle$. As we do not know what is the correct circuit to apply, we enumerate over all possible circuits of the blue shape (we can do it because its size is small). There are $2^{\mathcal{O}(d^2)}$ such circuits in total, and for each circuit we apply it to $|\psi\rangle$ and test if the state on A actually equals to $|0\rangle_A$ (we can do it by measuring many copies, and seeing the outcome all-0 with high probability). For now we assume that all local inversion circuits can be found exactly; this is addressed in more detail later.

At the end of this procedure, we end up with a list of candidate circuits $\mathcal{C}_A$ of the blue shape, such that each of them is a valid local inversion of A, i.e., for all $V_A \in \mathcal{C}_A$ we have $V_A|\psi\rangle = |0\rangle_A \otimes |\psi'\rangle$. The inverse of the lightcone of A in the unknown circuit U is among them, but we don't know which one. We repeat the same procedure for each region A, B, C, ... and get a list of candidate local inversions $\mathcal{C}_A$, $\mathcal{C}_B$, $\mathcal{C}_C$, ... for each region.

Note that in this construction shown in Fig. 3, only the local inversions acting on neighboring regions could overlap. For example, the blue and green circuit does not overlap because A and C are separated by distance $3d$, and each circuit could "spread" into region B for distance at most d.

The next observation is that there are certain blue circuits in $\mathcal{C}_A$ that share the same overlapping region with certain red circuits in $\mathcal{C}_B$, i.e. they share the same gates in the overlapping triangle of blue and red. For example, the inverse of the lightcone of A in U and the inverse of the lightcone of B in U share the same overlap. We call such circuits "consistent" with each other. Note that if two circuits are consistent, they can be merged into a bigger one. For example, take a blue circuit and a red circuit that are consistent, then they can be merged by considering the union of the gates, and applying the merged circuit to $|\psi\rangle$ will simultaneously invert both regions A and B. If we can find a local inversion for each region such that all nearest neighbors are consistent, then they can be merged into a depth-d circuit V that satisfies $V|\psi\rangle = |0^n\rangle$.

Now the task can be viewed as a constraint satisfaction problem: for each region, find a local inversion circuit among all candidate local inversions (there are at most $2^{\mathcal{O}(d^2)}$ choices), such that each pair of nearest neighbor circuits are consistent. This can be solved efficiently by a simple dynamic programming algorithm in time $n \cdot 2^{\mathcal{O}(d^2)}$.

To be more specific, suppose the system is divided into $L = \frac{n}{3d}$ regions of size $3d$ as in Fig. 3, and suppose we have found at most $M = 2^{\mathcal{O}(d^2)}$ local inversions for each region. These circuits are stored in an array C, where $C[i][j]$ denotes the jth local inversion circuit for the ith region. Define

an arrays *cost*, where $cost[i][j] = 0$ if there exists a consistent assignment at locations $1, 2, \ldots, i$ where $C[i][j]$ is used at location i; and $cost[i][j] \geq 1$ otherwise (let $cost[0][j] = 0$ for all j). Also define an array *prev*, where $prev[i][j]$ is an index k, such that there exists a consistent assignment at locations $1, 2, \ldots, i$ where $C[i][j]$ is used at location i and $C[i-1][k]$ is used at location $i-1$. $prev[i][j]$ is not defined when $cost[i][j] \geq 1$.

Once these arrays are constructed, we can take any circuit j such that $cost[L][j] = 0$, and construct a consistent assignment by tracing back through the *prev* array. Let *temp* be an array of size M. The following pseudocode shows how to construct these arrays in time $\mathcal{O}(LM^2)$.

```
for i = 1, 2, ..., L do
    for j = 1, 2, ..., M do
        for k = 1, 2, ..., M do
            temp[k] = cost[i − 1][k] + 1 [C[i][j] is not consistent with C[i − 1][k]]
        cost[i][j] = min_k temp[k]
        if cost[i][j] = 0 then
            prev[i][j] = arg min_k temp[k]
```

Finally, note that the above procedure can be implemented by a two-step process:

1. Learn reduced density matrices of $|\psi\rangle$ supported on the lightcone of each small region $A, B, C, \ldots$.
2. Find local inversions classically using the learned classical descriptions of the reduced density matrices, and then solve the constraint satisfaction problem.

This is because to find local inversions, say for the B region, we only need access to the reduced density matrix of $|\psi\rangle$ on the lightcone of B, which has $5d$ qubits, since the local inversion only acts on the reduced density matrix.

We need to learn $\frac{n}{3d}$ reduced density matrices of size at most $5d$. The following general lemma shows the complexity for learning reduced density matrices which we use throughout this section.

Lemma 23 (Learning reduced density matrices)**.** *Let ρ be an unknown n-qubit mixed state. Suppose we would like to learn its reduced density matrices $\rho_{A_1}, \ldots, \rho_{A_m}$ where A_i are subsystems of size at most k. Given a randomized measurement dataset $\mathcal{T}_\rho(N)$ of size $N = \frac{2^{\mathcal{O}(k)}}{\varepsilon^2} \log \frac{m}{\delta}$, we can learn a list of Hermitian matrices (not necessarily density matrices) $\{\sigma_{A_i}\}$ such that with probability at least $1 - \delta$, we have $\|\rho_{A_i} - \sigma_{A_i}\|_1 \leq \varepsilon$ for all i.*

Proof. Fix some i, we can write $\rho_{A_i} = \sum_{P \in \{I,X,Y,Z\}^{|A_i|}} \alpha_P P$. It suffices to learn the Pauli coefficients $\alpha_P = \frac{1}{2^{|A_i|}} \operatorname{Tr}(\rho_{A_i} P) = \frac{1}{2^{|A_i|}} \operatorname{Tr}(\rho P)$. Suppose we have learned these coefficients (denote as $\{\beta_P\}$) to within ε_1 precision. Let $\sigma_{A_i} := \sum_{P \in \{I,X,Y,Z\}^{|A_i|}} \beta_P P$, then

$$\|\rho_{A_i} - \sigma_{A_i}\|_1^2 \leq 2^{|A_i|} \operatorname{Tr}(\rho - \sigma)^2 = 2^{2|A_i|} \sum_P (\alpha_P - \beta_P)^2 \leq 2^{4k} \varepsilon_1^2, \tag{283}$$

which gives $\|\rho_{A_i} - \sigma_{A_i}\|_1 \leq 2^{2k} \varepsilon_1$. Thus to achieve $\|\rho_{A_i} - \sigma_{A_i}\|_1 \leq \varepsilon$ it suffices to learn $\{\operatorname{Tr}(\rho P)\}$ within accuracy $\varepsilon / 2^k$; there are at most $m \cdot 4^k$ k-local Pauli operators that we need to learn.

By the main result of [84], given a randomized measurement dataset of size

$$N = \frac{2^{\mathcal{O}(k)}}{\varepsilon^2} \log \frac{m}{\delta}, \tag{284}$$

with probability at least $1 - \delta$, we can learn all observables $\operatorname{Tr}(\rho P)$ for the $m \cdot 4^k$ k-local Pauli operators within accuracy $\varepsilon / 2^k$; this is sufficient to obtain Hermitian matrices $\{\sigma_{A_i}\}$ that satisfy $\|\rho_{A_i} - \sigma_{A_i}\|_1 \leq \varepsilon$ for all i. □

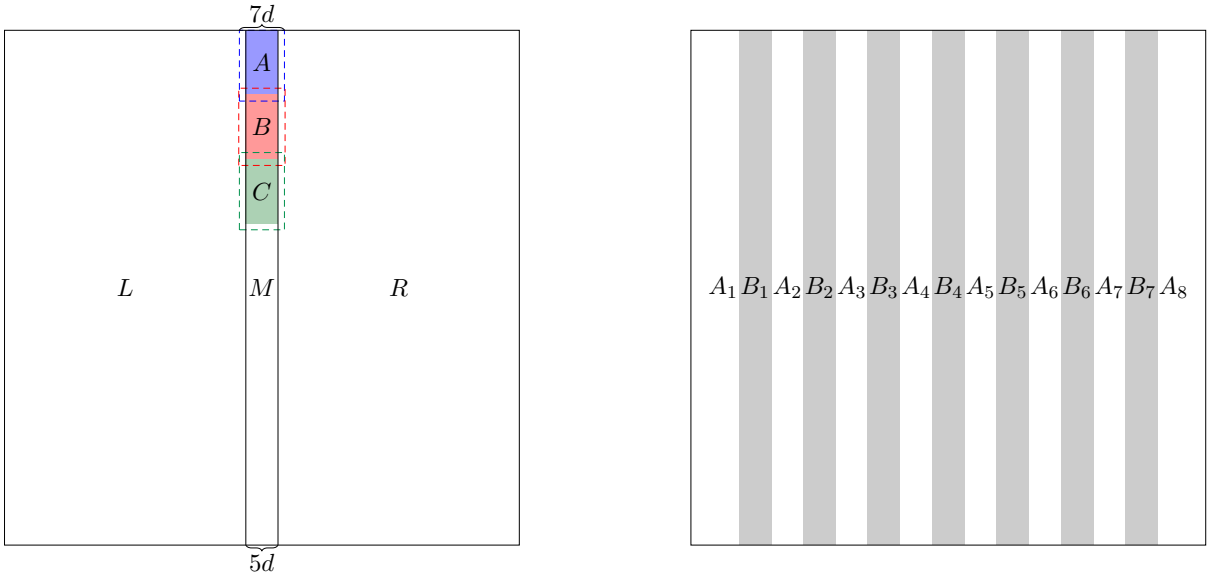


Figure 4: Learning to disentangle a quantum state generated by a shallow circuit in 2D. (a) The middle region M can be inverted by solving a similar 1D constraint satisfaction problem as in Fig. 3. (b) After inverting all the gray B_i regions, the remaining white A_i regions are disentangled into a tensor product of pure states.

Note that when the gates in the unknown circuit are assumed to come from a constant-size gate set, the reduced density matrices only have $2^{\mathcal{O}(d^2)} = \mathcal{O}(1)$ choices. Therefore, choosing ε to be some small constant in Lemma 23 suffices to learn all the reduced density matrices *exactly*. This allows us to find the exact local inversions by classically processing the reduced density matrices.

In summary, we have shown an algorithm that learns a depth-d circuit V that satisfies $|\psi\rangle = V^\dagger |0^n\rangle$ with success probability $1 - \delta$, using a randomized measurement dataset of size $N = \mathcal{O}(\log(n/\delta))$, in time $\mathcal{O}(n)$.

8.2 Disentangling a 2D state

Next we use the 1D techniques developed above to disentangle a state $|\psi\rangle = U |0^n\rangle$, where U is a depth-d circuit acting on a 2D lattice, for some constant $d = \mathcal{O}(1)$.

For this purpose we need to introduce a general property for quantum states generated by low depth circuits, that is they have finite correlation length.

Lemma 24 (Finite correlation length)**.** *Let $|\psi\rangle$ be a state generated by a depth-d geometrically-local circuit (Definition 9). Let A, B be two regions that are separated by distance at least $2d$ in the connectivity graph. Then $I(A:B)_\psi = 0$. In other words, let ρ_{AB}, ρ_A, ρ_B be the reduced density matrices of $|\psi\rangle$ on AB, A and B, then $\rho_{AB} = \rho_A \otimes \rho_B$.*

Proof. As A and B are separated by distance $2d$, their lightcones $L(A)$ and $L(B)$ are disjoint. $\rho_{AB} = \rho_A \otimes \rho_B$ follows from the fact that ρ_{AB} is generated by the gates in $L(AB)$, which is a tensor product between $L(A)$ and $L(B)$. □

Fig. 4 (a) shows a quantum state $|\psi\rangle$ (let $\rho = |\psi\rangle\langle\psi|$) prepared by a depth-$d$ circuit on a 2D lattice, divided into three regions L, M, R. Since L and R are separated by distance $5d$, Lemma 24

implies that $\rho_{LR} = \rho_L \otimes \rho_R$. Although subsystems L and R are not entangled with each other, they both could be entangled with M. Therefore we develop an argument to invert the qubits in M, so that the state on L and R could become a tensor product of pure states.

Note that M is a 1D-like region. Our goal is to find a depth-d circuit V acting on a slightly wider strip (of width $7d$) around M, such that $V|\psi\rangle = |0\rangle_M \otimes |\psi'\rangle$. Such a circuit exists since we can undo the lightcone of M, and we can find such a circuit using the same argument as in the previous section. In Fig. 4 (a), the blue, red and green regions play the same role as in Fig. 3. For example, we can find a set of local inversions $\mathcal{C}_A$ for the shaded blue region A, by first learning the reduced density matrix on the dotted blue region, and then enumerating over all depth-d circuits acting on the dotted blue region. After learning a set of local inversions for each local region, we can find a desired depth-d circuit that inverts M by solving a 1D constraint satisfaction problem.

Now, we have effectively reduced the problem of learning $|\psi\rangle$ to the following problem: given copies of a state $|\psi_1\rangle$ with the promise that

1. it is prepared by a depth-$2d$ circuit (defined on a 2D lattice) acting on $|0^n\rangle$;
2. its reduced density matrix on M equals $|0\rangle\langle 0|_M$.

The goal is to learn the state $|\psi_1\rangle$. Note that in this new state $\sigma = |\psi_1\rangle\langle\psi_1|$, even though its circuit depth has increased from d to $2d$, the reduced state on L and R is still in tensor product, i.e. $\sigma_{LR} = \sigma_L \otimes \sigma_R$, due to the fact that M (with width $5d$) is sufficiently wide. The main purpose of inverting the M region is that now σ_L and σ_R are guaranteed to be pure states, as shown by the following.

Lemma 25. *Let ρ_{ABC} be a pure state such that the following two properties hold:*

1. $\rho_B = |0\rangle\langle 0|_B$,
2. $\rho_{AC} = \rho_A \otimes \rho_C$.

Then ρ_A and ρ_C are both pure states.

Proof. This is a special case of Lemma 29. □

Next, we apply the above argument across the entire system. In Fig. 4 (b), the system is divided into many vertical strips of width $5d$. By repeating the above argument, we can learn a inverting circuit V_i for each shaded B_i region. Note that each V_i acts on a width-$7d$ strip around B_i and therefore different V_is do not overlap. By combining these different inverting circuits, overall we have learned a depth-d circuit V such that $V|\psi\rangle = |0\rangle_B \otimes |\psi'\rangle$ where B denotes the union of B_i.

Finally, by repeatedly applying Lemma 25, we know that the reduced density matrix of $V|\psi\rangle$ on each region A_i is a pure state. This means that overall the state can be written as $V|\psi\rangle = |0\rangle_B \otimes (\otimes_i |\phi\rangle_{A_i})$ for some pure states $|\phi\rangle_{A_i}$.

Now, we have disentangled the state $|\psi\rangle$ into a tensor product of many 1D-like pure states, and the problem of learning $|\psi\rangle$ is reduced to the following problem:

Problem 1. We are given copies of a state $|\psi_2\rangle$ with the promise that

1. it is prepared by a depth-$2d$ circuit (defined on a 2D lattice) acting on $|0^n\rangle$;
2. its reduced density matrix on each of the B_i regions in Fig. 4 (b) equals $|0\rangle\langle 0|_{B_i}$; in particular, this implies that $|\psi_2\rangle = |0\rangle_B \otimes (\otimes_i |\phi\rangle_{A_i})$ for some pure states $|\phi\rangle_{A_i}$.

The goal is to learn the state $|\psi_2\rangle$, and it suffices to learn each of the individual states $|\phi\rangle_{A_i}$.

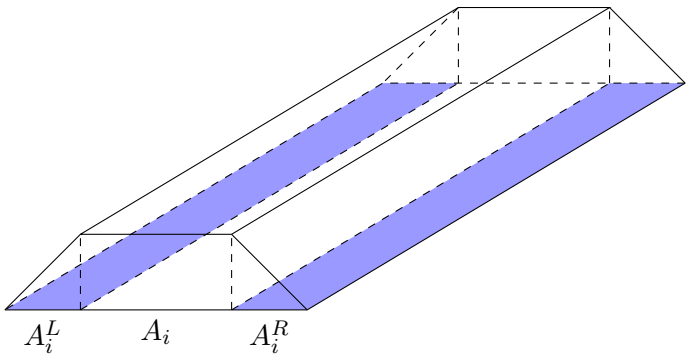


Figure 5: Each of the states on the white A_i regions in Fig. 4 (b) can be viewed as being prepared by a depth-$2d$ circuit acting on A_i (white) as well as ancilla qubits A_i^L and A_i^R (blue).

8.3 Learning finite correlated states in 1D

Next we show how to learn a state $|\phi\rangle$ (abbreviating the subscript A_i) on a specific region A_i that came from Problem 1. Besides the fact that $|\phi\rangle$ is a pure state, the learning algorithm heavily relies on the property that $|\phi\rangle$ is part of a larger state that is prepared by a depth-$2d$ circuit. Note that this does not imply that $|\phi\rangle$ itself can be prepared by a depth-$2d$ circuit acting on A_i. Instead, we will use this property to derive useful facts about $|\phi\rangle$, presented as two different viewpoints. Each of them leads to a learning algorithm that is similar to the approach in Section 8.1.

Viewpoint 1. By Lemma 24, the state $|\phi\rangle$ is a finite correlated state with correlation length $\ell = 4d$. That is, let $\sigma = |\phi\rangle\langle\phi|$ and let $R_1, R_2 \subseteq A_i$ be two regions that are separated by distance at least $4d$, then $\sigma_{R_1R_2} = \sigma_{R_1} \otimes \sigma_{R_2}$.

Viewpoint 2. $|\phi\rangle$ can be prepared by a depth-$2d$ circuit acting on A_i as well as some ancilla qubits A_i^L and A_i^R, shown in Fig. 5. To see this, recall that $|\phi\rangle$ is part of a state that is prepared by a depth-$2d$ circuit. Now, imagine that we *undo* all the gates in that circuit, except for those in the *backward lightcone* of A_i. This procedure does not affect the state on A_i, and the resulting circuit (denote as W_i) has exactly the same shape as in Fig. 5, where A_i^L, A_i^R both has width $2d$. Moreover, since $|\phi\rangle$ is a pure state, it is disentangled with the ancilla qubits, which means

$$W_i |0\rangle_{A_i^L} |0\rangle_{A_i} |0\rangle_{A_i^R} = |\text{junk}\rangle_{A_i^L} \otimes |\phi\rangle \otimes |\text{junk}'\rangle_{A_i^R} . \tag{285}$$

Clearly, Viewpoint 2 is a much stronger characterization of $|\phi\rangle$ and derives Viewpoint 1 as a corollary; however, it involves additional ancilla qubits. In the following, we show that each of these Viewpoints itself is sufficient to derive a learning algorithm; in particular,

- Using Viewpoint 1, we show that the state $|\phi\rangle$ can be prepared by a depth-$2^{\mathcal{O}(d^2)}$ circuit acting on A_i (without ancilla), therefore it can be learned using the techniques in Section 8.1.
- Using Viewpoint 2, we show how to learn a depth-$2d$ circuit W_i that prepares the state $|\phi\rangle$ using ancilla qubits, according to Eq. (285).

Central to both of these results is a technique that allows us to disentangle a finite correlated state in 1D. For simplicity, below we present this technique for a 1D system on a line with no width.

Lemma 26 (Disentangling finite correlated states in 1D)**.** *Let $|\phi\rangle$ be a state defined on a line with correlation length ℓ, that is, every two regions R_1, R_2 that are separated by distance at least ℓ*

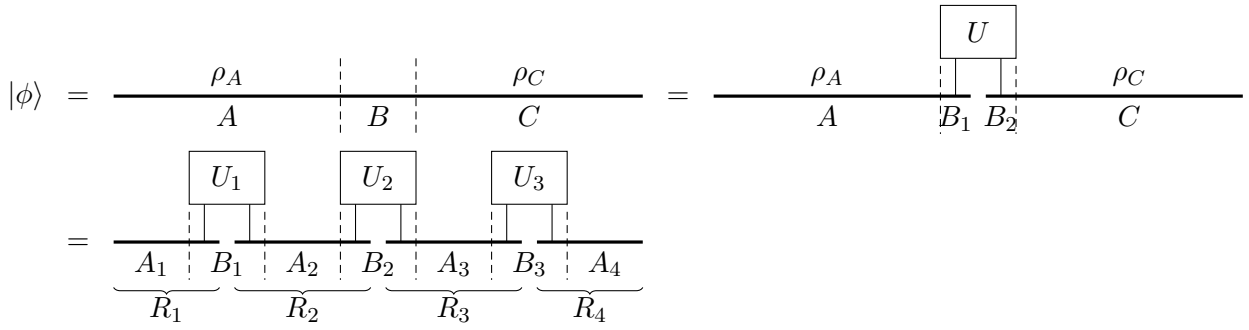


Figure 6: Disentangling a finite correlated state in 1D.

have zero mutual information, i.e. $\rho_{R_1R_2} = \rho_{R_1} \otimes \rho_{R_2}$*, where* $\rho = |\phi\rangle\langle\phi|$*. Divide the 1D line into contiguous regions of size* ℓ*, denote as* $A_1, B_1, A_2, B_2, \ldots, B_{L-1}, A_L$ *(Fig. 6). Then for each* i *there exists a unitary* U_i *acting on the* B_i *region, such that* $\prod_{i=1}^{L-1} U_i |\phi\rangle$ *is a tensor product of* L *pure states.*

Proof. We start with three subsystems A, B, C (first line of Fig. 6), where B has size ℓ. Then we have

$$\operatorname{rank}(\rho_B) = \operatorname{rank}(\rho_{AC}) = \operatorname{rank}(\rho_A \otimes \rho_C) = \operatorname{rank}(\rho_A) \cdot \operatorname{rank}(\rho_C) \leq \dim(B). \tag{286}$$

Purifying the state ρ_A (ρ_C) requires an ancilla system with dimension $\operatorname{rank}(\rho_A)$ ($\operatorname{rank}(\rho_C)$). Therefore we can partition B into two systems B_1, B_2, such that there exists pure states $|\phi_1\rangle_{AB_1}$ and $|\phi_2\rangle_{B_2C}$, such that $|\phi_1\rangle_{AB_1}$ is a purification of ρ_A, and $|\phi_2\rangle_{B_2C}$ is a purification of ρ_C. This implies that $|\phi_1\rangle_{AB_1} \otimes |\phi_2\rangle_{B_2C}$ is a purification of ρ_{AC}. Since $|\phi\rangle_{ABC}$ is also a purification of ρ_{AC}, by Uhlmann's theorem there exists a unitary U_B such that $|\phi\rangle_{ABC} = U_B |\phi_1\rangle_{AB_1} \otimes |\phi_2\rangle_{B_2C}$.

Applying this argument independently at different B_i regions (bottom line of Fig. 6), we have that for each $i = 1, 2, \ldots, L-1$, there exists a partition of the system B_i as two systems B_i^L and B_i^R, as well as a unitary U_i acting on $B_i = B_i^L \cup B_i^R$, such that

$$|\phi\rangle = U_i |\phi_1\rangle_{A_1 \ldots B_i^L} \otimes |\phi_2\rangle_{B_i^R A_{i+1} \cdots A_L}, \tag{287}$$

or equivalently, $U_i^\dagger |\phi\rangle = |\phi_1\rangle_{A_1 \ldots B_i^L} \otimes |\phi_2\rangle_{B_i^R A_{i+1} \cdots A_L}$, for some pure states $|\phi_1\rangle$ and $|\phi_2\rangle$. Next, we relabel the systems according to

$$R_i := B_{i-1}^R \cup A_i \cup B_i^L. \tag{288}$$

Intuitively, after applying all $U_i^\dagger$s, the system must be disentangled across all the R_i regions. To prove this we use a simple argument based on the strong subadditivity of quantum entropy (Lemma 27).

Let $\sigma := \left(\prod_{i=1}^{L-1} U_i^\dagger\right) |\phi\rangle\langle\phi| \left(\prod_{i=1}^{L-1} U_i\right)$ be the final (pure) state. Fix some i, our goal is to prove that σ_{R_i} is pure, i.e., $S(\sigma_{R_i}) = 0$. The strong subadditivity of quantum entropy gives

$$S(\sigma_{R_i}) \leq S(\sigma_{R_1 \ldots R_i}) + S(\sigma_{R_i \ldots R_L}) - S(\sigma) = S(\sigma_{R_1 \ldots R_i}) + S(\sigma_{R_i \ldots R_L}). \tag{289}$$

Note that when calculating $S(\sigma_{R_1 \ldots R_i})$ we can *undo* all the unitaries $U_j^\dagger$ for $j < i$ due to the invariance of entropy under unitary. Then $S(\sigma_{R_1 \ldots R_i}) = 0$ immediately follows from Eq. (287), and a similar argument shows $S(\sigma_{R_i \ldots R_L}) = 0$, which concludes the proof. □

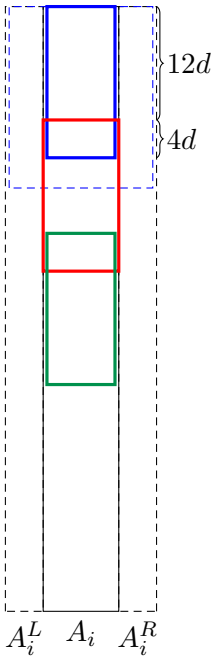


Figure 7: Learning a quantum state generated by a depth-$2d$ circuit with ancilla.

Lemma 27 (Strong subadditivity of quantum entropy [107]). *Let ρ be a mixed state defined on three systems A, B, C. Let $S(\rho) := -\operatorname{Tr}(\rho \log \rho)$ be the von Neumann entropy. Then we have*

$$S(\rho_{ABC}) + S(\rho_B) \leq S(\rho_{AB}) + S(\rho_{BC}). \tag{290}$$

Learning under Viewpoint 1. A corollary of Lemma 26 is that any finite correlated state in 1D can be prepared by a low-depth circuit, because each of the small pure state on the R_i regions in the bottom line of Fig. 6 can be prepared by a local unitary acting on $\mathcal{O}(\ell)$ qubits. Applying this argument to the state $|\phi\rangle_{A_i}$ shown in Fig. 5, we conclude that it can be prepared by two layers of unitaries acting on $\mathcal{O}(d^2)$ qubits, acting on the A_i region only. This implies that the state $|\phi\rangle_{A_i}$ can be prepared by a depth-$2^{\mathcal{O}(d^2)}$ circuit acting on A_i, and thus can be learned by applying the argument in Section 8.1.

Learning under Viewpoint 2. The main drawback of the above argument is that the learned circuit depth has an exponential blowup. To reduce this blowup we use additional structure of the state $|\phi\rangle_{A_i}$, described in Viewpoint 2 and Fig. 5. Note that there is a key difference between learning the state $|\phi\rangle_{A_i}$ and learning 1D states discussed in Section 8.1. Here, while the state $|\phi\rangle_{A_i}$ has a low-depth property shown in Fig. 5, this property relies on ancilla qubits (the $|\text{junk}\rangle$ states in Eq. (285)) that *we do not have access to*. Therefore we cannot directly apply the techniques in Section 8.1, which requires access to all qubits prepared by the low-depth circuit.

The main idea is to learn a mixed state ρ that is *locally consistent* with the state $|\phi\rangle\langle\phi|$, i.e., they have the same local reduced density matrices, and then show that this forces the two states to be globally the same.

The argument is illustrated in Fig. 7, where we learn to locally *prepare* the state instead of *invert* the state. Consider the state $|\phi\rangle$ on the A_i region shown in Fig. 7, and suppose we have learned its reduced density matrix ρ_{blue} on the solid blue region. Due to the fact that $|\phi\rangle$ is prepared by a depth-$2d$ circuit acting on A_i^L, A_i, A_i^R, we know that there exists a depth-$2d$ circuit acting on the

dotted blue region that prepares ρ_{blue} (the circuit looks like a small piece of Fig. 5), by undoing all the gates except for those in the backward lightcone of the solid blue region. We can perform a brute force search over all depth-$2d$ circuits acting on the dotted blue region, and for each of them we can test whether it prepares ρ_{blue}. In this way we obtain a list of depth-$2d$ circuits acting on the dotted blue region that prepares ρ_{blue}.

By repeating the above procedure we can obtain a list of local preparation circuits for each of the solid colored regions. A key point here is that the neighboring colored regions overlap by distance $4d$. Moreover, the local preparation circuits for the blue and green regions do not overlap, since the red region is sufficiently big. This enables us to solve a constraint satisfaction problem of the same nature as in Section 8.1, where we can choose a local preparation circuit for each region, such that neighboring circuits are consistent and can be merged together. Overall we have learned a depth-$2d$ circuit W acting on A_i^L, A_i, A_i^R, that simultaneously prepares all the local reduced density matrices.

Let $\rho := \mathrm{Tr}_{A_i^L A_i^R}(W\,|0\rangle\langle 0|_{A_i^L A_i A_i^R} W^\dagger)$ be the learned density matrix on A_i. At this point we know that ρ and $|\phi\rangle\langle\phi|$ are locally the same on the solid blue, red, and green regions (and so on), but this does not directly imply that $\rho = |\phi\rangle\langle\phi|$. For example, a Haar random pure state and the maximally mixed state are locally very close but globally very far. Next, we show that the finite correlation property forces ρ and $|\phi\rangle\langle\phi|$ to be globally equal.

Lemma 28 (Local consistency implies global consistency)**.** *Let $|\psi\rangle$ be a state defined on a 1D line with correlation length ℓ and let $\sigma = |\psi\rangle\langle\psi|$. Suppose the system is partitioned into contiguous regions $A_1, \ldots, A_L$ where $|A_i| \geq \ell$. Suppose ρ is a mixed state that satisfies $\rho_{A_i A_{i+1}} = \sigma_{A_i A_{i+1}}$ for all i, then $\rho = \sigma$.*

Proof. We show this for 3 subsystems; generalizing to more subsystems is straightforward. Let ρ be a mixed state satisfying $\rho_{A_1A_2} = \sigma_{A_1A_2}$ and $\rho_{A_2A_3} = \sigma_{A_2A_3}$. Following the proof of Lemma 26, there exists a unitary U acting on A_2 such that

$$U_{A_2}\,|\psi\rangle_{A_1A_2A_3} = |\phi_1\rangle_{A_1A_{21}} \otimes |\phi_2\rangle_{A_{22}A_3}\,, \tag{291}$$

where A_{21}, A_{22} is a partition of A_2, and $|\phi_1\rangle_{A_1A_{21}}$, $|\phi_2\rangle_{A_{22}A_3}$ are some pure states. Equivalently, we have

$$U_{A_2}\sigma U_{A_2}^\dagger = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}} \otimes |\phi_2\rangle\langle\phi_2|_{A_{22}A_3}\,. \tag{292}$$

Let $\tau := U_{A_2}\rho U_{A_2}^\dagger$, we will show that $\tau = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}} \otimes |\phi_2\rangle\langle\phi_2|_{A_{22}A_3}$, which implies $\rho = \sigma$.

First, taking the partial trace over A_3 on both sides of Eq. (292), we have

$$U\sigma_{A_1A_2}U^\dagger = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}} \otimes \mathrm{Tr}_{A_3}\,|\phi_2\rangle\langle\phi_2|\,. \tag{293}$$

Then, notice that

$$\tau_{A_1A_2} = U\rho_{A_1A_2}U^\dagger = U\sigma_{A_1A_2}U^\dagger = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}} \otimes \mathrm{Tr}_{A_3}\,|\phi_2\rangle\langle\phi_2|\,. \tag{294}$$

Tracing out A_{22} on both sides, we have $\tau_{A_1A_{21}} = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}}$; similarly, $\tau_{A_{22}A_3} = |\phi_2\rangle\langle\phi_2|_{A_{22}A_3}$. Since $\tau_{A_1A_{21}}$ and $\tau_{A_{22}A_3}$ are both pure states, this implies that the global state τ is a tensor product

$$\tau = \tau_{A_1A_{21}} \otimes \tau_{A_{22}A_3} = |\phi_1\rangle\langle\phi_1|_{A_1A_{21}} \otimes |\phi_2\rangle\langle\phi_2|_{A_{22}A_3}\,. \tag{295}$$

Thus we have $\tau = U\sigma U^\dagger$, which implies $\rho = \sigma$. □

Summary of our progress so far. So far we have developed all technical ingredients for learning a quantum state $|\psi\rangle = U|0^n\rangle$, under the simplified setting that U is a depth $d = \mathcal{O}(1)$ circuit acting on a 2D lattice, and each gate in U is from a constant size gate set.

Note that all the above arguments can be viewed as first learning the local reduced density matrices of $|\psi\rangle$ followed by classically reconstructing the circuit. As we have discussed before in Section 8.1, a reduced density matrix of constant size can be learned *exactly* as it only has a constant number of choices. In the disentangling step shown in Fig. 4, we can learn $\mathcal{O}(n)$ reduced density matrices on the dotted regions of size $\mathcal{O}(d^2)$, and then classically reconstruct a depth-d circuit V in time $\mathcal{O}(n)$, such that $V|\psi\rangle = |0\rangle_B \otimes (\otimes_i |\phi\rangle_{A_i})$ where the pure states $|\phi\rangle_{A_i}$ live on the white regions of Fig. 4 (b).

Proof of second claim of Theorem 9. Next, we start with Viewpoint 2. As shown in Fig. 7, learning a state $|\phi\rangle_{A_i}$ requires learning its reduced density matrices of size $5d \times 16d$. This can be achieved by experimentally applying V to $|\psi\rangle$ and then learning the reduced density matrices. Equivalently, say we want to learn the reduced density matrix of $|\phi\rangle_{A_i}$ on a region M of size $5d \times 16d$, then it suffices to learn a reduced density matrix of $|\psi\rangle$ of size $7d \times 18d$ on a region surrounding M, then *classically* apply the gates of V within the backward lightcone of M, and then classically trace out the qubits outside M. In other words, the reduced density matrices of $|\phi\rangle_{A_i}$ can be simulated by slightly larger reduced density matrices of $|\psi\rangle$. Using these reduced density matrices, for each i we can learn a depth-$2d$ circuit W_i such that

$$W_i|0\rangle_{A_i^L A_i A_i^R} = |\phi\rangle_{A_i} \otimes |\text{junk}\rangle_{A_i^L A_i^R}, \tag{296}$$

which takes total time $\mathcal{O}(n)$. The entire process requires $\mathcal{O}(n)$ reduced density matrices of $|\psi\rangle$ of size $\mathcal{O}(d^2)$, which can be learned exactly with probability at least $1-\delta$, using a randomized measurement dataset of size $N = \mathcal{O}(\log(n/\delta))$.

The state $|\psi\rangle$ can be prepared as follows:

1. Initialize registers A_i, B_i, A_i^L, A_i^R in the state $|0\rangle$. Let $A = \cup_i A_i$ and $B = \cup_i B_i$.
2. For each i, apply the depth-$2d$ circuit W_i to $A_i^L A_i A_i^R$.
3. Apply the depth-d circuit $V^\dagger$ to AB, and the state $|\psi\rangle$ lives on AB.

Overall the learned circuit has depth $3d$ and can be implemented on an extended 2D lattice, where the qubits in A_i can interact with its ancilla qubits A_i^L, A_i^R as well as neighboring B_i regions.

In Fig. 7 we have chosen the width of A_i to be $5d$. Note that the width of A_i^L and A_i^R are both $2d$, regardless of the width of A_i. In fact we could have chosen the width of A_i to be Cd for some large constant C, and the number of ancilla qubits is at most $n/(Cd) \cdot 4d = \frac{4}{C}n$, which can be made arbitrarily small.

Proof of third claim of Theorem 9. Using Viewpoint 1, the state $|\phi\rangle_{A_i}$ can be prepared by a depth-$2^{\mathcal{O}(d^2)}$ circuit acting on A_i, and thus can be learned by applying the argument in Section 8.1. Let $|\phi\rangle_{A_i} = W|0\rangle_{A_i}$ for some depth-$2^{\mathcal{O}(d^2)}$ circuit W acting on A_i. A technical issue here is that we no longer have the guarantee that W consists of gates from a finite gate set as in U, because the existence of W comes from the disentangling argument in Lemma 26, instead of coming from the original circuit U as in Viewpoint 2. Below we discuss how to find this circuit W.

Let $d' = 2^{\mathcal{O}(d^2)}$ be the circuit depth of W. Following Section 8.1, we can learn reduced density matrices of $\sigma := |\phi\rangle\langle\phi|_{A_i}$ of size $5d \times 5d'$ (which can be done exactly, as discussed above) and then

classically find local inversions for regions of size $5d \times 3d'$. Following Fig. 3, let A be a region of size $5d \times 3d'$, and let AA_1 be the lightcone of A with size $5d \times 4d'$. Then there is a depth-d' circuit W_{AA_1} acting on AA_1 such that

$$\mathrm{Tr}_{A_1}\left(W_{AA_1}\sigma_{AA_1}W_{AA_1}^\dagger\right) = |0\rangle\langle 0|_A . \tag{297}$$

To find the local inversion W_{AA_1} we use an ε_0-net over depth-d' circuits acting on AA_1, denoted as $\mathcal{N}_{\varepsilon_0}(AA_1)$ (see Definition 18 and Lemma 19), which has size at most

$$S = \left(\frac{d'^3}{\varepsilon_0}\right)^{\mathcal{O}(d'^3)} . \tag{298}$$

By definition, there exists $\hat{W}_{AA_1} \in \mathcal{N}_{\varepsilon_0}(AA_1)$ such that $\|\hat{W}_{AA_1} - W_{AA_1}\|_\infty \le \varepsilon_0$, which gives

$$\langle 0_A|\mathrm{Tr}_{A_1}\left(\hat{W}_{AA_1}\sigma_{AA_1}\hat{W}_{AA_1}^\dagger\right)|0_A\rangle \ge 1 - 2\varepsilon_0. \tag{299}$$

By enumerating over every element in $\mathcal{N}_{\varepsilon_0}(AA_1)$, we can find a list of circuits which satisfy the above equation. Following the argument in Section 8.1, we repeat the same procedure for each local region and merge the local circuits into a global depth-d' circuit $\hat{W}_i$, which approximately inverts each local region up to $1 - 2\varepsilon_0$ fidelity. By union bound, we have

$$\left|\langle 0_{A_i}|\,\hat{W}_i\,|\phi\rangle_{A_i}\right|^2 \ge 1 - 2\sqrt{n}\varepsilon_0. \tag{300}$$

After learning each region A_i, the state $|\psi\rangle$ can be approximately prepared as follows:

1. Initialize registers A_i, B_i in the state $|0\rangle$. Let $A = \cup_i A_i$ and $B = \cup_i B_i$.
2. For each i, apply the depth-d' circuit $\hat{W}_i^\dagger$ to A_i.
3. Apply the depth-d circuit $V^\dagger$ to AB, and the state on AB, which is $|\hat{\psi}\rangle = V^\dagger(\otimes_i \hat{W}_i^\dagger)\,|0^n\rangle$, approximately equals to $|\psi\rangle$.

We bound the approximation error as follows.

$$\left|\langle\hat{\psi}|\psi\rangle\right|^2 = \left|\langle 0^n|\,(\otimes_i \hat{W}_i)V\,|\psi\rangle\right|^2 = \prod_i \left|\langle 0_{A_i}|\,\hat{W}_i\,|\phi\rangle_{A_i}\right|^2 \ge 1 - 2n\varepsilon_0. \tag{301}$$

Therefore to achieve $1-\varepsilon$ fidelity it suffices to choose $\varepsilon_0 = \frac{\varepsilon}{2n}$, which gives total running time $n \cdot S = (n/\varepsilon)^{\mathcal{O}(1)}$.

8.4 Robustness to imprecision

In the previous sections we have been focusing on a finite gateset, which allows us to learn reduced density matrices exactly, and therefore the disentangling procedure in Fig. 4 can be performed exactly. However, it's not clear that this argument still works for general SU(4) gates, because in this case each step can only be performed *approximately*. In particular, we can only approximately disentangle the state using the procedure in Fig. 4, and learning the remaining 1D states poses new technical challenges as they are no longer pure.

In this section we address this issue. In the following we first outline the argument and develop key technical lemmas, before going into the full proof of the first claim in Theorem 9.

We start with the disentangling step in Fig. 4. Here, instead of exhaustively enumerating small circuits acting on local regions, we can only enumerate over an ε-net of the circuit. Therefore, we are only able to find circuits that approximately invert each B_i region shown in Fig. 4 (b). This means that after the disentangling step, the reduced density matrix on B will be *close* to $|0\rangle\langle 0|_B$, instead of being *exactly equal* to $|0\rangle\langle 0|_B$.

Now the question is what happens to the remaining A_i regions. Note that the state is still in tensor product across different A_i regions due to the finite correlation length property, but the reduced density matrices on each A_i region will not be pure. The following lemma shows that these states are *approximately pure*.

Lemma 29. *Let $\rho_{A_1A_2\ldots A_LB}$ be a pure state such that the following two properties hold:*

1. $\langle 0_B|\rho_B|0_B\rangle \geq 1-\varepsilon$,
2. $\rho_{A_1A_2\ldots A_L} = \rho_{A_1}\otimes\cdots\otimes\rho_{A_L}$.

Then for each $i = 1,\ldots,L$ there exists a pure state $|\phi\rangle_{A_i}$ such that $\langle\phi_{A_i}|\rho_{A_i}|\phi_{A_i}\rangle \geq 1-\varepsilon$.

Proof. Consider the operator norm $\|\rho\|_\infty := \lambda_{\max}(\rho) = \max_{|\psi\rangle}\langle\psi|\rho|\psi\rangle$. Condition 1 gives $\|\rho_B\|_\infty \geq 1-\varepsilon$. Using condition 2 we have

$$\|\rho_B\|_\infty = \|\rho_{A_1\ldots A_L}\|_\infty = \|\rho_{A_1}\otimes\cdots\otimes\rho_{A_L}\|_\infty = \prod_{i=1}^{L}\|\rho_{A_i}\|_\infty \geq 1-\varepsilon, \tag{302}$$

which implies that $\lambda_{\max}(\rho_{A_i}) \geq 1-\varepsilon$ for any i. □

Next, we discuss how to learn these states $\{\rho_{A_i}\}$ that are approximately pure. Again, we still have the property that each ρ_{A_i} is a 1D-like state with finite correlation length. However, our previous techniques developed in Section 8.3 only work for *exactly* pure states. We develop new techniques by examining the robustness of the key technical lemma developed in Section 8.3, Lemma 26.

There are two key ingredients in the proof of Lemma 26:

1. The use of Uhlmann's theorem to prove the existence of a local disentangling unitary;
2. The use to entropy inequalities (in particular, strong subadditivity) to prove that the state is disentangled into many local pieces after applying Uhlmann's unitaries across the entire system.

Fortunately, both ingredients are robust. First, Uhlmann's theorem says that if two mixed states are close, then there exists a unitary (acting on the purifying system) that approximately maps between their purifications. Second, entropy inequalities are robust, thanks to the continuity of entropy given below.

Lemma 30 (Fannes–Audenaert inequality)**.** *Let ρ, σ be two n-qubit density matrices, and let $\varepsilon := \frac{1}{2}\|\rho-\sigma\|_1$. Then*

$$|S(\rho) - S(\sigma)| \leq n\varepsilon + h(\varepsilon), \tag{303}$$

where $h(\cdot)$ is the binary entropy function and can be upper bounded as $h(\varepsilon) \leq 2\sqrt{\varepsilon}$.

We formalize the above intuitions as the following main technical lemma, which is a robust version of Lemma 28.

Lemma 31. *Let ρ be an n-qubit mixed state defined on systems $A_1, \ldots, A_L$, with the following properties:*

1. *there exists an n-qubit pure state $|\psi\rangle$, such that $\langle\psi|\rho|\psi\rangle \geq 1-\varepsilon$.*
2. *for any $i = 2, 3, \ldots, L-1$, it holds that $I(A_1 \cdots A_{i-1} : A_{i+1} \cdots A_L)_\rho = 0$.*

For simplicity we assume that L is odd. Let σ be another n-qubit mixed state that satisfies

$$\frac{1}{2}\|\sigma_{A_{2i}A_{2i+1}A_{2i+2}} - \rho_{A_{2i}A_{2i+1}A_{2i+2}}\|_1 \leq \delta, \quad \forall i = 0, 1, \ldots, (L-1)/2, \tag{304}$$

Then

$$\frac{1}{2}\|\sigma - \rho\|_1 \leq 13n\varepsilon^{1/16} + 4n\delta^{1/4}. \tag{305}$$

Proof. The above condition says that ρ and σ are close on local regions $A_1A_2, A_2A_3A_4, A_4A_5A_6, \ldots, A_{L-1}A_L$. The goal is to prove that they are globally close.

Let $\tau := |\psi\rangle\langle\psi|$ denote the density matrix of $|\psi\rangle$. For any $j \in \{1, 2, \ldots, (L-1)/2\}$, define three regions $L^{(j)} := A_{\leq 2j-1}$, $M^{(j)} := A_{2j}$, $R^{(j)} := A_{\geq 2j+1}$ (the superscript (j) is abbreviated when there is no confusion).

Note that for any subsystem W, we have

$$\frac{1}{2}\|\tau_W - \rho_W\|_1 \leq \frac{1}{2}\|\tau - \rho\|_1 \leq \sqrt{1 - \langle\psi|\rho|\psi\rangle} \leq \sqrt{\varepsilon}. \tag{306}$$

Therefore,

$$\begin{aligned}\|\tau_{LR} - \tau_L \otimes \tau_R\|_1 &\leq \|\tau_{LR} - \rho_{LR}\|_1 + \|\rho_{LR} - \rho_L \otimes \rho_R\|_1 + \|\rho_L \otimes \rho_R - \tau_L \otimes \tau_R\|_1 \\ &\leq \|\tau_{LR} - \rho_{LR}\|_1 + \|\rho_L - \tau_L\|_1 + \|\rho_R - \tau_R\|_1 \\ &\leq \varepsilon_1\end{aligned} \tag{307}$$

where we let $\varepsilon_1 := 6\sqrt{\varepsilon}$. Then, the relationship between fidelity and trace distance implies that

$$F(\tau_{LR}, \tau_L \otimes \tau_R) \geq 1 - \|\tau_{LR} - \tau_L \otimes \tau_R\|_1 \geq 1 - \varepsilon_1. \tag{308}$$

Let $|\phi_1\rangle_{LM_1^{(j)}}$ be a purification of τ_L, and let $|\phi_2\rangle_{M_2^{(j)}R}$ be a purification of τ_R. Note that $\dim(M_1^{(j)}) \leq \dim(L)$ and $\dim(M_2^{(j)}) \leq \dim(R)$. Let $M'^{(j)}$ be an ancilla space with dimension $\dim(M_1^{(j)})\dim(M_2^{(j)})/\dim(M^{(j)})$. Here $M'^{(j)}$ is needed in case $M^{(j)}$ is smaller than $M_1^{(j)}M_2^{(j)}$. Now, $|\psi\rangle_{LMR}|0\rangle_{M'^{(j)}}$ is a purification of the state τ_{LR}, while $|\phi_1\rangle_{LM_1^{(j)}} \otimes |\phi_2\rangle_{M_2^{(j)}R}$ is a purification of the state $\tau_L \otimes \tau_R$, and they have the same dimension. Then by Uhlmann's theorem, there exists a unitary $U^{(j)} : M^{(j)}M'^{(j)} \to M_1^{(j)}M_2^{(j)}$, such that

$$U^{(j)}_{M^{(j)}M'^{(j)}} |\psi\rangle_{LM^{(j)}R} |0\rangle_{M'^{(j)}} \approx_{\varepsilon_1} |\phi_1\rangle_{LM_1^{(j)}} \otimes |\phi_2\rangle_{M_2^{(j)}R}. \tag{309}$$

Here, $|u\rangle \approx_\varepsilon |v\rangle$ means $|\langle u|v\rangle|^2 \geq 1-\varepsilon$.

The above argument shows the existence of a unitary $U^{(j)}$ acting on $M^{(j)} = A_{2j}$ (as well as an ancilla system $M'^{(j)}$), that approximately disentangles the state $|\psi\rangle$ into a tensor product between $LM_1^{(j)}$ and $M_2^{(j)}R$, where $M_1^{(j)}$, $M_2^{(j)}$ are ancilla systems associated with A_{2j}. We apply all such unitaries $U^{(j)}$ ($j \in \{1, 2, \ldots, (L-1)/2\}$) to $|\psi\rangle$, and obtain

$$\eta := \left(\prod_{j=1}^{(L-1)/2} U^{(j)}\right) |\psi\rangle\langle\psi| \otimes |0\rangle\langle 0|_{M'} \left(\prod_{j=1}^{(L-1)/2} U^{(j)\dagger}\right), \tag{310}$$

where M' represents the union of all $M'^{(j)}$. Note that η supports on $A_1, A_3, A_5, \ldots, A_L$ as well as $M_1^{(j)}, M_2^{(j)}$ for $j \in \{1, 2, \ldots, (L-1)/2\}$. Now, we relabel the systems according to

$$B_j := M_2^{(j-1)} \cup A_{2j-1} \cup M_1^{(j)}, \quad j \in \{1, 2, \ldots, (L+1)/2\}, \tag{311}$$

and the state η supports on B_j, $j \in \{1, 2, \ldots, (L+1)/2\}$, and we want to prove that it is approximately a tensor product across all B_j regions via upper bounding the relative entropy

$$D(\eta || \otimes_j \eta_{B_j}) = \sum_j S(\eta_{B_j}) - S(\eta) = \sum_j S(\eta_{B_j}). \tag{312}$$

By the strong subadditivity of quantum entropy,

$$S(\eta_{B_j}) \leq S(\eta_{B_{\leq j}}) + S(\eta_{B_{\geq j}}) - S(\eta) = S(\eta_{B_{\leq j}}) + S(\eta_{B_{\geq j}}). \tag{313}$$

Focusing on the entropy of $S(\eta_{B_{\leq j}})$, we can ignore the unitaries that are applied on regions other than A_{2j}. Note that Eq. (309) implies that

$$\frac{1}{2} \left\| \mathrm{Tr}_{M_2^{(j)} R} (U^{(j)} |\psi\rangle\langle\psi| \otimes |0\rangle\langle 0|_{M'^{(j)}} U^{(j)\dagger}) - |\phi_1\rangle\langle\phi_1|_{L M_1^{(j)}} \right\|_1 \leq \sqrt{\varepsilon_1}. \tag{314}$$

Therefore by the Fannes-Audenaert inequality,

$$S(\eta_{B_{\leq j}}) = S(\mathrm{Tr}_{M_2^{(j)} R} (U^{(j)} |\psi\rangle\langle\psi| \otimes |0\rangle\langle 0|_{M'^{(j)}} U^{(j)\dagger})) \leq 2|L|\sqrt{\varepsilon_1} + 2\varepsilon_1^{1/4} \leq 2n\sqrt{\varepsilon_1} + 2\varepsilon_1^{1/4}. \tag{315}$$

A similar argument holds for $S(\eta_{B_{\geq j}})$. Therefore we have

$$S(\eta_{B_j}) \leq 4n\sqrt{\varepsilon_1} + 4\varepsilon_1^{1/4}, \quad \forall j \in \{1, 2, \ldots, (L+1)/2\}. \tag{316}$$

Let

$$\omega := \left(\prod_{j=1}^{(L-1)/2} U^{(j)} \right) \sigma \otimes |0\rangle\langle 0|_{M'} \left(\prod_{j=1}^{(L-1)/2} U^{(j)\dagger} \right), \tag{317}$$

then $\|\sigma - |\psi\rangle\langle\psi|\|_1 = \|\omega - \eta\|_1$. Note that for any j, η_{B_j} only depends on the reduced density matrix $\tau_{A_{2j-2} A_{2j-1} A_{2j}}$; similarly, ω_{B_j} only depends on the reduced density matrix $\sigma_{A_{2j-2} A_{2j-1} A_{2j}}$. Therefore,

$$\begin{aligned} \left\| \omega_{B_j} - \eta_{B_j} \right\|_1 &\leq \left\| \sigma_{A_{2j-2} A_{2j-1} A_{2j}} - \tau_{A_{2j-2} A_{2j-1} A_{2j}} \right\|_1 \\ &\leq \left\| \sigma_{A_{2j-2} A_{2j-1} A_{2j}} - \rho_{A_{2j-2} A_{2j-1} A_{2j}} \right\|_1 + \left\| \rho_{A_{2j-2} A_{2j-1} A_{2j}} - \tau_{A_{2j-2} A_{2j-1} A_{2j}} \right\|_1 \\ &\leq 2\delta + 2\sqrt{\varepsilon}. \end{aligned} \tag{318}$$

Note that $|B_j| \leq 3n$, by the Fannes-Audenaert inequality,

$$S(\omega_{B_j}) \leq S(\eta_{B_j}) + 3n(\delta + \sqrt{\varepsilon}) + 2\sqrt{\delta + \sqrt{\varepsilon}}. \tag{319}$$

This implies that

$$\begin{aligned} D(\omega || \otimes_j \omega_{B_j}) &= \sum_j S(\omega_{B_j}) - S(\omega) \\ &\leq \sum_j S(\omega_{B_j}) \\ &\leq \sum_j S(\eta_{B_j}) + 3n^2(\delta + \sqrt{\varepsilon}) + 2n\sqrt{\delta + \sqrt{\varepsilon}}. \end{aligned} \tag{320}$$

Then

$$\begin{aligned}
\|\sigma-\rho\|_1 &\le \|\sigma-\tau\|_1 + \|\tau-\rho\|_1 \\
&\le \|\omega-\eta\|_1 + 2\sqrt{\varepsilon} \\
&\le \left\|\omega-\otimes_j\omega_{B_j}\right\|_1 + \left\|\otimes_j\omega_{B_j}-\otimes_j\eta_{B_j}\right\|_1 + \left\|\otimes_j\eta_{B_j}-\eta\right\|_1 + 2\sqrt{\varepsilon} \\
&\le \sqrt{2D(\omega||\otimes_j\omega_{B_j})} + 2n\delta + 2n\sqrt{\varepsilon} + \sqrt{2D(\eta||\otimes_j\eta_{B_j})} + 2\sqrt{\varepsilon} \\
&\le \sqrt{8n(n\sqrt{\varepsilon_1}+\varepsilon_1^{1/4}) + 6n^2(\delta+\sqrt{\varepsilon}) + 4n\sqrt{\delta+\sqrt{\varepsilon}}} \\
&\quad + \sqrt{8n(n\sqrt{\varepsilon_1}+\varepsilon_1^{1/4})} + 2n\delta + 2(n+1)\sqrt{\varepsilon}.
\end{aligned} \tag{321}$$

Here in the fourth line we use the quantum Pinsker inequality, which says that $\|\rho-\sigma\|_1 \le \sqrt{2D(\rho\|\sigma)}$ for two density matrices ρ,σ. Using the fact that $\varepsilon_1 = 6\sqrt{\varepsilon}$, we have

$$\begin{aligned}
\frac{1}{2}\|\sigma-\rho\|_1 &\le n\delta + 2n\sqrt{\varepsilon} + \sqrt{8}n\varepsilon_1^{1/4} + \sqrt{8}\sqrt{n}\varepsilon_1^{1/8} + \frac{\sqrt{6}}{2}n\sqrt{\delta} + \frac{\sqrt{6}}{2}n\varepsilon^{1/4} + \sqrt{n}\delta^{1/4} + \sqrt{n}\varepsilon^{1/8} \\
&\le \sqrt{8}n\varepsilon_1^{1/4} + \sqrt{8}\sqrt{n}\varepsilon_1^{1/8} + 5n\varepsilon^{1/8} + 4n\delta^{1/4} \\
&\le 13n\varepsilon^{1/16} + 4n\delta^{1/4}.
\end{aligned} \tag{322}$$

□

Finally, the next technical lemma bounds the distance between the learned state and the unknown state $|\psi\rangle$.

Lemma 32. *Let $|\psi\rangle_{A_1\ldots A_LB}$ be a pure state, and let $\rho_{A_1\ldots A_LB} = |\psi\rangle\langle\psi|_{A_1\ldots A_LB}$. Suppose the following two properties hold:*

1. $\langle 0_B|\rho_B|0_B\rangle = 1-\varepsilon$,
2. $\rho_{A_1\ldots A_L} = \rho_{A_1}\otimes\cdots\otimes\rho_{A_L}$.

Suppose $\{\sigma_{A_i}\}$ are density matrices that satisfies $\frac{1}{2}\|\rho_{A_i}-\sigma_{A_i}\|_1 \le \delta$ for any i. Then

$$\frac{1}{2}\left\|(\otimes_{i=1}^L\sigma_{A_i})\otimes|0\rangle\langle 0|_B - |\psi\rangle\langle\psi|\right\|_1 \le \sqrt{2\varepsilon+L\delta}. \tag{323}$$

Proof. The state $|\psi\rangle_{A_1\ldots A_LB}$ can be written as

$$|\psi\rangle_{A_1\ldots A_LB} = \sqrt{1-\varepsilon}\,|0\rangle_B\,|\phi\rangle_{A_1\ldots A_L} + \sqrt{\varepsilon}\,|\text{else}\rangle_{A_1\ldots A_LB}\,, \tag{324}$$

where $\langle 0|_B\,|\text{else}\rangle_{A_1\ldots A_LB} = 0$. This implies that

$$\rho_{A_1\ldots A_L} = \mathrm{Tr}_B\,\rho_{A_1\ldots A_LB} = (1-\varepsilon)\,|\phi\rangle\langle\phi|_{A_1\ldots A_L} + \varepsilon\,\mathrm{Tr}_B\,|\text{else}\rangle\langle\text{else}|\,. \tag{325}$$

Note that

$$\begin{aligned}
\frac{1}{2}\|\rho_{A_1\ldots A_L}-\sigma_{A_1}\otimes\cdots\otimes\sigma_{A_L}\|_1 &= \frac{1}{2}\|\rho_{A_1}\otimes\cdots\otimes\rho_{A_L}-\sigma_{A_1}\otimes\cdots\otimes\sigma_{A_L}\|_1 \\
&\le \frac{1}{2}\sum_{i=1}^L\|\rho_{A_i}-\sigma_{A_i}\|_1 \\
&\le L\delta.
\end{aligned} \tag{326}$$

Therefore,

$$\begin{aligned}\langle\psi|_{A_1\ldots A_L B}\,\sigma_{A_1}\otimes\cdots\otimes\sigma_{A_L}\otimes|0\rangle\langle 0|_B\,|\psi\rangle_{A_1\ldots A_L B} &\geq \langle\psi|\,\rho_{A_1\ldots A_L}\otimes|0\rangle\langle 0|_B\,|\psi\rangle - L\delta\\ &\geq (1-\varepsilon)^2 - L\delta\\ &\geq 1-2\varepsilon-L\delta.\end{aligned} \tag{327}$$

This implies that

$$\frac{1}{2}\left\|(\otimes_{i=1}^{L}\sigma_{A_i})\otimes|0\rangle\langle 0|_B - |\psi\rangle\langle\psi|\right\|_1 \leq \sqrt{2\varepsilon+L\delta}. \tag{328}$$

□

Proof of first claim of Theorem 9. Next we show how to use the above techniques to learn an unknown quantum state $|\psi\rangle = U|0^n\rangle$, with the promise that U is a depth-d circuit acting on a 2D lattice (here d is treated as a generic parameter which is not necessarily a constant) with arbitrary SU(4) gates.

We work with Viewpoint 2 described in Section 8.3. As discussed at the end of Section 8.3, the learning process requires $\mathcal{O}(n)$ reduced density matrices of $|\psi\rangle$ of size $\mathcal{O}(d^2)$. Suppose all of these reduced density matrices are learned to within ε_0 trace distance with probability $1-\delta$, then by Lemma 23 it suffices to take a randomized measurement dataset $\mathcal{T}_{|\psi\rangle}(N)$ of size

$$N = \frac{2^{\mathcal{O}(d^2)}}{\varepsilon_0^2}\log\frac{n}{\delta}. \tag{329}$$

Next we proceed with the disentangling step shown in Fig. 4. We have learned the reduced density matrices on the dotted regions shown in Fig. 4 (a) to within ε_0 trace distance. Denote the dotted blue region as AA_1 where A is the colored blue region, and let ρ_{AA_1} be the reduced density matrix of $|\psi\rangle$ on AA_1. We know that there exists a depth-$2d$ circuit V_{AA_1} such that

$$V_{AA_1}\rho_{AA_1}V_{AA_1}^\dagger = |0\rangle\langle 0|_A\otimes\sigma_{A_1} \tag{330}$$

for some density matrix σ_{A_1}. We have learned a density matrix $\hat{\rho}_{AA_1}$ such that $\|\hat{\rho}_{AA_1}-\rho_{AA_1}\|_1\leq\varepsilon_0$. To find an approximate local inversion for the region A, we perform a brute force search over an ε_0-net for depth-$2d$ circuits acting on AA_1, denoted as $\mathcal{N}_{\varepsilon_0}(AA_1)$, which is constructed by discretizing each SU(4) gate (see Definition 18 and Lemma 19), which has size at most

$$S = \left(\frac{d^3}{\varepsilon_0}\right)^{\mathcal{O}(d^3)}. \tag{331}$$

Note that Eq. (330) together with $\|\hat{\rho}_{AA_1}-\rho_{AA_1}\|_1\leq\varepsilon_0$ implies that

$$\mathrm{Tr}\Big(\langle 0|_A\,V_{AA_1}\hat{\rho}_{AA_1}V_{AA_1}^\dagger\,|0\rangle_A\Big)\geq 1-\varepsilon_0. \tag{332}$$

By definition of ε_0-net, there exists a unitary $\hat{V}_{AA_1}\in\mathcal{N}_{\varepsilon_0}(AA_1)$ that satisfies $\|\hat{V}_{AA_1}-V_{AA_1}\|_\infty\leq\varepsilon_0$, which gives

$$\mathrm{Tr}\Big(\langle 0|_A\,\hat{V}_{AA_1}\hat{\rho}_{AA_1}\hat{V}_{AA_1}^\dagger\,|0\rangle_A\Big)\geq 1-2\varepsilon_0. \tag{333}$$

The algorithm is to enumerate over all elements in $\mathcal{N}_{\varepsilon_0}(AA_1)$ and find the ones which satisfy the above equation. Each of these circuits is an approximate local inversion in the sense that

$$\mathrm{Tr}\Big(\langle 0|_A\,\hat{V}_{AA_1}\rho_{AA_1}\hat{V}_{AA_1}^\dagger\,|0\rangle_A\Big)\geq 1-3\varepsilon_0. \tag{334}$$

Using the same argument as in Section 8.2, in Fig. 4 (a) we can find a depth-d circuit $\hat{V}$ acting on the width-$7d$ strip around M, such that Eq. (334) is satisfied for all local colored regions. There are at most $\sqrt{n}$ such regions. Let $\rho = |\psi\rangle\langle\psi|$, by union bound,

$$\mathrm{Tr}\Big(\langle 0|_M \hat{V}\rho\hat{V}^\dagger |0\rangle_M\Big) \geq 1 - 3\sqrt{n}\varepsilon_0. \tag{335}$$

Repeat the same procedure for all vertical B_i strips shown in Fig. 4 (b). There are at most $\sqrt{n}$ different vertical strips. Let $B = \cup_i B_i$, and let V denote the union of all learned inversion circuits across different regions, we have

$$\mathrm{Tr}\Big(\langle 0|_B V\rho V^\dagger |0\rangle_B\Big) \geq 1 - 3n\varepsilon_0. \tag{336}$$

Now, the problem reduces to learning the state $V|\psi\rangle$, which can be formulated as follows.

Problem 2. We are given copies of a state $\sigma = |\phi\rangle\langle\phi|$ with the promise that

1. it is prepared by a depth-$2d$ circuit (defined on a 2D lattice) acting on $|0^n\rangle$;
2. its reduced density matrix on each of the B_i regions in Fig. 4 (b) is close $|0\rangle\langle 0|_{B_i}$, i.e. $\langle 0_B|\sigma_B|0_B\rangle \geq 1 - \varepsilon_1$.

The goal is to (approximately) learn the state $|\phi\rangle$.

Let $|\phi\rangle := V|\psi\rangle$ and let $\varepsilon_1 := 3n\varepsilon_0$. Consider dividing the state $\sigma = |\phi\rangle\langle\phi|$ into regions $A_1, A_2, \ldots, A_L$ and $B = \cup_i B_i$ as in Fig. 4 (b). As the regions $\{A_i\}$ are sufficiently far from each other, the reduced density matrix on $A = \cup_i A_i$ is a tensor product across each region, i.e., $\sigma_{A_1\ldots A_L} = \sigma_{A_1} \otimes \cdots \otimes \sigma_{A_L}$. By Eq. (336), we have $\langle 0_B|\sigma_B|0_B\rangle \geq 1 - \varepsilon_1$. By Lemma 29, for each $i = 1, \ldots, L$ there exists a pure state $|\phi\rangle_{A_i}$ such that $\langle\phi_{A_i}|\sigma_{A_i}|\phi_{A_i}\rangle \geq 1 - \varepsilon_1$.

Next we discuss how to learn the state σ_{A_i} for a fixed i. This is similar to the earlier situation in Viewpoint 2, but with the critical difference that here σ_{A_i} is no longer pure. So we list the updated Viewpoint below.

Viewpoint 2'. σ_{A_i} can be prepared by a depth-$2d$ circuit acting on A_i as well as some ancilla qubits A_i^L and A_i^R, shown in Fig. 5. To see this, recall that σ_{A_i} is part of a state that is prepared by a depth-$2d$ circuit. Now, imagine that we *undo* all the gates in that circuit, except for those in the *backward lightcone* of A_i. This procedure does not affect the state on A_i, and the resulting circuit (denote as W_i) has exactly the same shape as in Fig. 5, where A_i^L, A_i^R both has width $2d$. Note that here σ_{A_i} could be entangled with the ancilla qubits, and we have

$$\mathrm{Tr}_{A_i^L A_i^R}\Big(W_i|0\rangle\langle 0|_{A_i^L A_i A_i^R} W_i^\dagger\Big) = \sigma_{A_i}. \tag{337}$$

Using the same argument as the end of Section 8.3, the reduced density matrices of σ_{A_i} can be simulated by reduced density matrices of $|\psi\rangle\langle\psi|$ on slightly larger regions. Therefore we can obtain reduced density matrices of σ_{A_i} within trace distance ε_0. Let C be the solid blue region in Fig. 7, and let CC_1 be the dotted blue region. We have learned a reduced density matrix $\hat{\sigma}_C$ such that $\|\hat{\sigma}_C - \sigma_C\|_1 \leq \varepsilon_0$. From Viewpoint 2', we know that there is a depth-$2d$ circuit W_{CC_1} acting on CC_1, such that

$$\mathrm{Tr}_{C_1}\Big(W_{CC_1}|0\rangle\langle 0|_{CC_1} W_{CC_1}^\dagger\Big) = \sigma_C. \tag{338}$$

Consider an ε_0-net for depth-$2d$ circuits acting on CC_1, denoted as $\mathcal{N}_{\varepsilon_0}(CC_1)$. By definition, there exists a unitary $\hat{W}_{CC_1}$ that satisfies $\|\hat{W}_{CC_1} - W_{CC_1}\|_\infty \leq \varepsilon_0$, which means that

$$\begin{aligned}
&\left\|\mathrm{Tr}_{C_1}\left(\hat{W}_{CC_1}|0\rangle\langle 0|_{CC_1}\hat{W}_{CC_1}^\dagger\right) - \hat{\sigma}_C\right\|_1 \\
\leq &\left\|\mathrm{Tr}_{C_1}\left(\hat{W}_{CC_1}|0\rangle\langle 0|_{CC_1}\hat{W}_{CC_1}^\dagger\right) - \sigma_C\right\|_1 + \|\sigma_C - \hat{\sigma}_C\|_1 \\
\leq &2\varepsilon_0.
\end{aligned} \tag{339}$$

By enumerating over every element in $\mathcal{N}_{\varepsilon_0}(CC_1)$, we can find a list of circuits $\{\hat{W}'_{CC_1}\}$ that satisfy $\left\|\mathrm{Tr}_{C_1}\left(\hat{W}'_{CC_1}|0\rangle\langle 0|_{CC_1}\hat{W}'^\dagger_{CC_1}\right) - \hat{\sigma}_C\right\|_1 \leq 2\varepsilon_0$. Any such circuit $\hat{W}'_{CC_1}$ will also satisfy

$$\left\|\mathrm{Tr}_{C_1}\left(\hat{W}'_{CC_1}|0\rangle\langle 0|_{CC_1}\hat{W}'^\dagger_{CC_1}\right) - \sigma_C\right\|_1 \leq 3\varepsilon_0. \tag{340}$$

Using the same argument as in Section 8.3, we can merge these learned local circuits into a global depth-$2d$ circuit $\hat{W}_i$. Let $\hat{\sigma}_{A_i} := \mathrm{Tr}_{A_i^L A_i^R}\left(\hat{W}_i|0\rangle\langle 0|_{A_i^L A_i A_i^R}\hat{W}_i^\dagger\right)$ be the learned reduced density matrix on A_i, then the local reduced density matrices of $\hat{\sigma}_{A_i}$ and σ_{A_i} are $3\varepsilon_0$ close in trace distance on solid colored regions in Fig. 7. This allows us to invoke the main technical lemma, Lemma 31, which gives

$$\frac{1}{2}\|\hat{\sigma}_{A_i} - \sigma_{A_i}\|_1 \leq 13n\varepsilon_1^{1/16} + 8n\varepsilon_0^{1/4} \leq 22n^{17/16}\varepsilon_0^{1/16}. \tag{341}$$

The state $|\psi\rangle$ can be approximately prepared as follows:

1. Initialize registers A_i, B_i, A_i^L, A_i^R in the state $|0\rangle$. Let $A = \cup_i A_i$ and $B = \cup_i B_i$.
2. For each i, apply the depth-$2d$ circuit $\hat{W}_i$ to $A_i^L A_i A_i^R$. The reduced density matrix on AB equals $(\otimes_i \hat{\sigma}_{A_i}) \otimes |0\rangle\langle 0|_B$
3. Apply the depth-d circuit $V^\dagger$ to AB, and the reduced density matrix on AB is $\hat{\rho} = V^\dagger(\otimes_i \hat{\sigma}_{A_i})\otimes |0\rangle\langle 0|_B V$, which approximately equals to $|\psi\rangle\langle\psi|$.

Similar to the proof of second claim of Theorem 9 at the end of Section 8.3, we can choose the A_i regions to be sufficiently wide, such that the number of ancilla qubits equals to tn for an arbitrarily small constant t.

The final task is to bound the error between the learned density matrix and $|\psi\rangle\langle\psi|$. Using Lemma 32, the trace distance can be bounded as

$$\begin{aligned}
\frac{1}{2}\left\|V^\dagger(\otimes_i \hat{\sigma}_{A_i})\otimes|0\rangle\langle 0|_B V - |\psi\rangle\langle\psi|\right\|_1 &= \frac{1}{2}\left\|(\otimes_i \hat{\sigma}_{A_i})\otimes|0\rangle\langle 0|_B - V|\psi\rangle\langle\psi|V^\dagger\right\|_1 \\
&\leq \sqrt{2\cdot 3n\varepsilon_0 + \sqrt{n}\cdot 22n^{17/16}\varepsilon_0^{1/16}} \\
&\leq 6n^{25/32}\varepsilon_0^{1/32}.
\end{aligned} \tag{342}$$

Therefore, to achieve trace distance ε, it suffices to choose $\varepsilon_0 = \mathcal{O}(\frac{\varepsilon^{32}}{n^{25}})$. The total sample complexity is

$$N = \frac{2^{\mathcal{O}(d^2)}}{\varepsilon_0^2}\log\frac{n}{\delta} = \frac{2^{\mathcal{O}(d^2)}n^{50}}{\varepsilon^{64}}\log\frac{n}{\delta}. \tag{343}$$

The total running time is

$$n\cdot S = \left(\frac{nd^3}{\varepsilon}\right)^{\mathcal{O}(d^3)}. \tag{344}$$

9 Verifying learned shallow circuits under average-case distance

From the previous appendices, we have seen that given an n-qubit CPTP map $\mathcal{C}$ promised to be a unitary U generated by a constant-depth quantum circuit, we can learn a constant-depth $2n$-qubit circuit $\hat{V}$, such that $\hat{V}$ is close to $U \otimes U^\dagger$, and the reduced channel $\hat{\mathcal{E}} := \mathcal{E}^{\hat{V}}_{\leq n}$ of $\hat{V}$ on the first n qubits is close to $\mathcal{C} = U(\cdot)U^\dagger = \mathcal{U}$ in the diamond distance. In this section, we answer the question: What happens if there is no promise that $\mathcal{C}$ is a unitary generated by a shallow quantum circuit, and, furthermore, $\mathcal{C}$ may not even be unitary?

Given an arbitrary CPTP map $\mathcal{C}$, the proposed algorithm can still learn a constant-depth $2n$-qubit circuit $\hat{V}$ with an associated n-qubit CPTP map $\hat{\mathcal{E}} := \mathcal{E}^{\hat{V}}_{\leq n}$. However, without the promise on $\mathcal{C}$, the learned map $\hat{\mathcal{E}}$ could be arbitrary. This raises the question: can we verify that $\hat{\mathcal{E}}$ is close to $\mathcal{C}$? From the previous section on the hardness for learning log-depth circuits, we see that even if $\mathcal{C}$ is an n-qubit unitary U generated by a log-depth circuit, one already needs $\exp(\Omega(n))$ queries to check if U is close to I in the diamond distance or not. Hence, when the learning algorithm outputs $\hat{\mathcal{E}} = \mathcal{I}$, which is very likely in this case as the unitary U_x in Eq. (277) is almost identity, we cannot efficiently check if $\hat{\mathcal{E}}$ is close to $\mathcal{C}$ in the diamond distance. The exponential hardness stems from the definition of diamond distance, which considers the worst case over all possible input states.

To circumvent the exponential hardness, we consider closeness under the average-case distance $\mathcal{D}_{\text{ave}}$ (see Definition 3) instead of the worst-case distance $\mathcal{D}_\diamond$. We give a verification algorithm that verifies the learned map $\hat{\mathcal{E}}$ by outputting PASS or FAIL as follows:

1. the verification algorithm outputs FAIL with high probability if the learned map $\hat{\mathcal{E}}$ is not close to $\mathcal{C}$ under the average-case distance $\mathcal{D}_{\text{ave}}$;

2. the verification algorithm outputs PASS with high probability if the learned map $\hat{\mathcal{E}}$ is close to $\mathcal{C}$ under the average-case distance $\mathcal{D}_{\text{ave}}$ and the unknown map $\mathcal{C}$ is close to a unitary.

The verification algorithm only needs access to a randomized measurement dataset $\mathcal{T}_{\mathcal{C}}(N)$ generalizing Definition 8 by replacing the unitary U with the map $\mathcal{C}$. Formally, we have the following theorem.

Theorem 10 (Verifying the learned shallow circuit)**.** *Given a failure probability δ, a verification error ε, a learned constant-depth $2n$-qubit circuit $\hat{V}$, the associated n-qubit CPTP map $\hat{\mathcal{E}} = \mathcal{E}^{\hat{V}}_{\leq n}$, and an unknown n-qubit CPTP map $\mathcal{C}$. With a randomized measurement dataset $\mathcal{T}_{\mathcal{C}}(N)$ of size*

$$N = \mathcal{O}\left(\frac{n^2 \log(n/\delta)}{\varepsilon^2}\right), \tag{345}$$

the verification algorithm outputs PASS or FAIL such that

1. *if $\mathcal{D}_{\text{ave}}(\hat{\mathcal{E}}, \mathcal{C}) > \varepsilon$, the output is FAIL with probability $\geq 1 - \delta$.*

2. *if $\mathcal{D}_{\text{ave}}(\hat{\mathcal{E}}, \mathcal{C}) \leq \frac{\varepsilon}{12n}$ and $\left\|\mathcal{C}^\dagger \mathcal{C} - \mathcal{I}\right\|_\diamond \leq \frac{\varepsilon}{12n}$, the output is PASS with probability $\geq 1 - \delta$;*

The computational time of the verification algorithm is $\mathcal{O}(nN)$.

Proof. The verification algorithm is based on the concept of weak approximate local identity presented in Section 4.2. Let us define the n-qubit CPTP map

$$\hat{\mathcal{I}} := \hat{\mathcal{E}}^\dagger \mathcal{C}. \tag{346}$$

Note that $\hat{\mathcal{E}}^\dagger(\rho)$ can be implemented by appending n-qubit maximally mixed state to ρ, evolving $\rho \otimes (I_n/2^n)$ under the unitary $\hat{V}^\dagger$, then tracing out the appended n ancilla qubits, i.e.,

$$\hat{\mathcal{E}}^\dagger(\rho) = \mathrm{Tr}_{>n}\left(\hat{V}^\dagger(\rho \otimes I_n/2^n)\hat{V}\right), \tag{347}$$

where I_n is an n-qubit identity. The verification algorithm uses the randomized measurement dataset $\mathcal{T}_{\mathcal{C}}(N)$ to estimate $\hat{o}_i$ approximating $\mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}}, \mathcal{I})$ up to $\varepsilon/(3n)$ error for all i from 1 to n with probability at least $1-\delta$. Then the verification algorithm outputs

$$\begin{cases} \textsc{pass}, & \text{if } \frac{3}{2}\sum_{i=1}^n \hat{o}_i \leq \varepsilon/2, \\ \textsc{fail}, & \text{if } \frac{3}{2}\sum_{i=1}^n \hat{o}_i > \varepsilon/2. \end{cases} \tag{348}$$

From Lemma 33 presented at the end of this section, we can show that the dataset size N stated in Eq. (345) is sufficient to guarantee the desired property on $\hat{o}_i$ and the computational time to estimate $\hat{o}_i$ for all i is $\mathcal{O}(nN)$. We define the event that

$$\left|\hat{o}_i - \mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}}, \mathcal{I})\right| \leq \frac{\varepsilon}{6n}, \quad \forall i = 1, \ldots, n \tag{349}$$

to be event E^*. Conditioning on event E^*, we show that the desired outputs, FAIL and PASS, must be given by the verification algorithm in the two scenarios stated in the theorem, respectively.

Case 1: $\mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{E}}, \mathcal{C}) > \varepsilon$. When conditioning on event E^*, we claim that the algorithm always outputs FAIL. We prove this claim by contradiction. Assume that the algorithm outputs PASS. From the definition of fidelity $F(\rho,\sigma) = \mathrm{Tr}\left(\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}}\right)^2$ given in Definition 2, we can see that $F(\rho,\sigma) \geq \mathrm{Tr}(\rho\sigma)$. Hence, from Definition 3 on $\mathcal{D}_{\mathrm{ave}}$, we have

$$\varepsilon < \mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{E}}, \mathcal{C}) \leq \mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{E}}^\dagger \mathcal{C}, \mathcal{I}) = \mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{I}}, \mathcal{I}). \tag{350}$$

If the algorithm outputs PASS, we have

$$\frac{3}{2}\sum_{i=1}^n \hat{o}_i \leq \frac{\varepsilon}{2}. \tag{351}$$

Because in the event E^*, Eq. (349) ensures

$$\left|\hat{o}_i - \mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}}, \mathcal{I})\right| \leq \frac{\varepsilon}{6n}, \tag{352}$$

we can conclude that

$$\frac{3}{2}\sum_{i=1}^n \mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}}, \mathcal{I}) \leq \frac{3}{4}\varepsilon. \tag{353}$$

Using Lemma 6 on global identity check from weak local identity check, we have

$$\mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{I}}, \mathcal{I}) \leq \frac{3}{2}\sum_{i=1}^n \mathcal{D}_{\mathrm{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}}, \mathcal{I}) \leq \frac{3}{4}\varepsilon. \tag{354}$$

This inequality contradicts the one in Eq. (350). Hence, if $\mathcal{D}_{\mathrm{ave}}(\hat{\mathcal{E}}, \mathcal{C}) > \varepsilon$, the output of the verification algorithm is FAIL with probability at least $1-\delta$.

Case 2: $\mathcal{D}_{\text{ave}}(\hat{\mathcal{E}},\mathcal{C}) \leq \varepsilon/(24n)$ and $\left\|\mathcal{C}^\dagger\mathcal{C}-\mathcal{I}\right\|_\diamond \leq \varepsilon/(12n)$. When conditioning on event E^*, we claim that the algorithm always outputs PASS. We begin by noting that the fidelity $F(\rho,\sigma) \leq F(\mathcal{E}(\rho),\mathcal{E}(\sigma))$ for any CPTP map $\mathcal{E}$ from Fact 1. Therefore, we have

$$\mathcal{D}_{\text{ave}}(\mathcal{C}^\dagger\hat{\mathcal{E}},\mathcal{C}^\dagger\mathcal{C}) \leq \mathcal{D}_{\text{ave}}(\hat{\mathcal{E}},\mathcal{C}) \leq \frac{\varepsilon}{24n}. \tag{355}$$

We now consider the following derivations,

$$\begin{aligned}
\mathcal{D}_{\text{ave}}(\hat{\mathcal{I}},\mathcal{I}) &= \mathcal{D}_{\text{ave}}(\hat{\mathcal{E}}^\dagger\mathcal{C},\mathcal{I}) && (356)\\
&= \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-\mathcal{F}((\hat{\mathcal{E}}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)\right] && (357)\\
&= \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-\text{Tr}\Big(\mathcal{C}(|\psi\rangle\langle\psi|)\hat{\mathcal{E}}(|\psi\rangle\langle\psi|)\Big)\right] && (358)\\
&= \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)\right]. && (359)
\end{aligned}$$

Using the triangle inequality for Fubini-Study metric Θ from Fact 1, we have

$$\begin{aligned}
&\sqrt{1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)} && (360)\\
&\leq \sin\left(\Theta\left((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|)\right)+\Theta\left((\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|\right)\right) && (361)\\
&\leq \sin\left(\Theta\left((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|)\right)\right)+\sin\left(\Theta\left((\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|\right)\right) && (362)\\
&\leq \sqrt{1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|))}+\sqrt{1-F((\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)}. && (363)
\end{aligned}$$

From $1-F(\rho,\psi) \leq \frac{1}{2}\|\rho-\psi\|_1$ for any state ρ and pure state ψ from Fact 1, we have

$$1-F((\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|) \leq \frac{1}{2}\left\|(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|)-|\psi\rangle\langle\psi|\right\|_{\text{tr}} \leq \frac{\varepsilon}{24n}. \tag{364}$$

From the two inequalities above, we see that

$$\sqrt{1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)} \leq \sqrt{1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|))}+\sqrt{\frac{\varepsilon}{24n}}. \tag{365}$$

Using Jensen's inequality, the above inequality, and Eq. (356), we obtain

$$\begin{aligned}
&\mathcal{D}_{\text{ave}}(\hat{I},I) && (366)\\
&= \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),|\psi\rangle\langle\psi|)\right] && (367)\\
&\leq \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|))\right]+\frac{\varepsilon}{24n} && (368)\\
&\quad+2\sqrt{\frac{\varepsilon}{24n}}\sqrt{\mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1-F((\mathcal{C}^\dagger\hat{\mathcal{E}})(|\psi\rangle\langle\psi|),(\mathcal{C}^\dagger\mathcal{C})(|\psi\rangle\langle\psi|))\right]} && (369)\\
&= \mathcal{D}_{\text{ave}}(\mathcal{C}^\dagger\hat{\mathcal{E}},\mathcal{C}^\dagger\mathcal{C})+\frac{\varepsilon}{24n}+2\sqrt{\frac{\varepsilon}{24n}}\sqrt{\mathcal{D}_{\text{ave}}(\mathcal{C}^\dagger\hat{\mathcal{E}},\mathcal{C}^\dagger\mathcal{C})} \leq \frac{\varepsilon}{6n}. && (370)
\end{aligned}$$

The last inequality follows from Eq. (355). Using Lemma 5 on weak local identity from global identity check through average-case distance, we have

$$\mathcal{D}_{\text{ave}}(\mathcal{E}_i^{\hat{\mathcal{I}}},\mathcal{I}) \leq \frac{\varepsilon}{6n} \tag{371}$$

for all i from 1 to n. When event E^* occurs, we can combine the above with Eq. (349) to show that

$$\hat{o}_i \leq \frac{\varepsilon}{3n}, \quad \forall i = 1, \ldots, n. \tag{372}$$

As a result, we can see that $\frac{3}{2}\sum_{i=1}^{n} \hat{o}_i \leq \varepsilon/2$. Hence, in this case, the output of the verification algorithm is PASS with probability at least $1 - \delta$. □

From the theorem, the verification algorithm outputs PASS with high probability if the promise on $\mathcal{C}$ is satisfied, and one uses our proposed learning algorithm to learn $\hat{\mathcal{E}}$. Furthermore, whenever the verification algorithm outputs PASS, we can be certain that $\hat{\mathcal{E}}$ is close to $\mathcal{C}$ (under the average-case distance). Together, our proposed learning algorithm and verification algorithm enable one to learn a verifiable shallow quantum circuit approximation to an arbitrary unknown CPTP map $\mathcal{C}$.

Lemma 33 (Checking weak approximate local identity)**.** *Given a failure probability δ, a verification error ε, a learned constant-depth $2n$-qubit circuit $\hat{V}$, the associated n-qubit CPTP map $\hat{\mathcal{E}} = \mathcal{E}^{\hat{V}}_{\leq n}$, and an unknown n-qubit CPTP map $\mathcal{C}$. With a randomized measurement dataset $\mathcal{T}_{\mathcal{C}}(N)$ of size*

$$N = \mathcal{O}\left(\frac{n^2 \log(n/\delta)}{\varepsilon^2}\right), \tag{373}$$

we can estimate $\hat{o}_i, \forall i$ in time $\mathcal{O}(nN)$ such that

$$\left|\hat{o}_i - \mathcal{D}_{\text{ave}}(\mathcal{E}_i^{\hat{\mathcal{E}}^\dagger \mathcal{C}}, \mathcal{I})\right| \leq \frac{\varepsilon}{3n}, \quad \forall i = 1, \ldots, n, \tag{374}$$

with probability at least $1 - \delta$.

Proof. Recall from Eq. (347) that the CPTP map $\hat{E}^\dagger$ is given by

$$\hat{\mathcal{E}}^\dagger(\rho) = \text{Tr}_{>n}\left(\hat{V}^\dagger(\rho \otimes I_n/2^n)\hat{V}\right). \tag{375}$$

Hence, we have the following identity for the single-qubit CPTP map,

$$\mathcal{E}_i^{\hat{\mathcal{E}}^\dagger \mathcal{C}}(\rho_i) = \text{Tr}_{\neq i}\left(\hat{V}^\dagger(\mathcal{C}(\rho_i \otimes I_{n-1}/2^{n-1}) \otimes I_n/2^n)\hat{V}\right), \tag{376}$$

where ρ_i is a single-qubit density matrix, $\rho_i \otimes I_{n-1}/2^{n-1}$ is an n-qubit density matrix equal to ρ_i on the i-th qubit and maximally mixed on all other qubits, and $\text{Tr}_{\neq i}$ traces out all qubits except for the i-th qubit. Because $\hat{V}$ is a constant-depth quantum circuit, $\mathcal{E}_i^{\hat{\mathcal{E}}^\dagger \mathcal{C}}$ depends only on a reduced channel $\mathcal{E}^{\mathcal{C}}_{S_i}$ of $\mathcal{C}$ on a subset S_i of qubits with $|S_i| = \mathcal{O}(1)$ and $i \in S_i$, i.e.,

$$\mathcal{E}_i^{\hat{\mathcal{E}}^\dagger \mathcal{C}}(\rho_i) = \text{Tr}_{\neq i}\left(\hat{V}^\dagger\left(\left(\mathcal{E}^{\mathcal{C}}_{S_i} \otimes \mathcal{I}_{[n]\setminus S_i}\right)(\rho_i \otimes I_{n-1}/2^{n-1}) \otimes I_n/2^n\right)\hat{V}\right), \tag{377}$$

where $\mathcal{I}_{[n]\setminus S_i}$ is the identity CPTP map over qubit 1 to qubit n not in set S_i. For any $i = 1, \ldots, n$, from the results in [85–87, 108], one could use $\mathcal{T}_{\mathcal{C}}(N)$ with the specified size to learn $\hat{\mathcal{E}}^{\mathcal{C}}_{S_i}$ such that

$$\left\|\hat{\mathcal{E}}^{\mathcal{C}}_{S_i} - \mathcal{E}^{\mathcal{C}}_{S_i}\right\|_\diamond \leq \frac{\varepsilon}{3n}, \tag{378}$$

with probability at least $1 - (\delta/n)$. By the union bound, we have

$$\left\|\hat{\mathcal{E}}^{\mathcal{C}}_{S_i} - \mathcal{E}^{\mathcal{C}}_{S_i}\right\|_\diamond \leq \frac{\varepsilon}{3n}, \quad \forall i = 1, \ldots, n, \tag{379}$$

with probability at least $1-\delta$. Hence, from Eq. (377), we can learn $\hat{\mathcal{E}}_i^{\hat{\mathcal{E}}^\dagger\mathcal{C}}$ for all i such that

$$\left\|\hat{\mathcal{E}}_i^{\hat{\mathcal{E}}^\dagger\mathcal{C}} - \mathcal{E}_i^{\hat{\mathcal{E}}^\dagger\mathcal{C}}\right\|_\diamond \leq \frac{\varepsilon}{3n}, \quad \forall i = 1,\ldots,n, \tag{380}$$

with probability at least $1-\delta$. By defining

$$\hat{o}_i := \mathcal{D}_{\text{ave}}\left(\hat{\mathcal{E}}_i^{\hat{\mathcal{E}}^\dagger\mathcal{C}}, \mathcal{I}\right) = \mathop{\mathbb{E}}_{|\psi\rangle:\text{Unif}}\left[1 - \langle\psi|\,\hat{\mathcal{E}}_i^{\hat{\mathcal{E}}^\dagger\mathcal{C}}(|\psi\rangle\langle\psi|)\,|\psi\rangle\right], \quad \forall i = 1,\ldots,n, \tag{381}$$

we can obtain the desired claim. □

10 Exponentially many local minima in parameterized shallow quantum circuits

In this section, we study the optimization landscape of training 1D shallow parameterized quantum circuits to learn an unknown unitary. In particular, we will show that there are exponentially many strictly suboptimal local minima, where each local minimum is the minimum over an exponentially sized neighborhood. Consider a simple 1D shallow parameterized quantum circuit,

$$U(\vec{\theta}) := \prod_j \exp(i\theta_{1,j}\,\text{SWAP}_{2j+1,2j+2}) \prod_j \exp(i\theta_{2,j}\,\text{SWAP}_{2j,2j+1}) \prod_j \exp(i\theta_{3,j}\,\text{SWAP}_{2j+1,2j+2}), \tag{382}$$

where $\vec{\theta} = (\theta_{1,j},\theta_{2,j},\theta_{3,j})$ is a vector of all the real-valued parameters. We consider an unknown unitary U over n qubits to be given by the tensor product of SWAP operators over some pairs of qubits, i.e.,

$$U_S = \prod_{i\in S} \text{SWAP}_{i,i+3}, \tag{383}$$

for some subset $S \subseteq \{0,1,2,\ldots,\lfloor n/4\rfloor - 1\}$ of qubits with $|S| = \Theta(n)$. For any such subset S, there exists a parameter vector $\vec{\theta}$ such that $U_S = U(\vec{\theta})$.

To avoid barren plateaus in the optimization landscape, we consider the local cost function [17],

$$C_S(\vec{\theta}) := \mathop{\mathbb{E}}_{|\psi\rangle=\bigotimes_{i=1}^n |\psi_i\rangle\in\text{stab}_1^{\otimes n}} \sum_{i=1}^n \left(1 - \text{Tr}\left(\langle\psi_i|\,U(\vec{\theta})^\dagger U_S\,|\psi\rangle\langle\psi|\,U_S^\dagger U(\vec{\theta})\,|\psi_i\rangle\right)\right) \geq 0. \tag{384}$$

It is well known that the local cost function is faithful [17, 109], i.e., if the local cost function is at most ε, then U is close to $U(\vec{\theta})$ up to average-case distance (equiv. to normalized Frobenius norm; See Prop. 1) of $\mathcal{O}(\varepsilon)$, and when U_S is ε-close to $U(\vec{\theta})$ in the average-case distance, the local cost function is bounded above by $\mathcal{O}(n\varepsilon)$. The local cost function does not suffer from the barren plateau problem when $U(\vec{\theta})$ and U_S can both be implemented by shallow quantum circuits. For those unfamiliar with barren plateau, it is an overwhelmingly large region in the parameter space with a large cost function and a near-zero gradient [17, 28]. When a barren plateau is present, one can easily randomly initialize on the barren plateau and cannot escape the plateau.

While no barren plateau is present in training shallow parameterized circuits, we show that there are exponentially many strictly suboptimal local minima in the optimization landscape. Furthermore, these suboptimal local minima are minima over neighborhoods with an exponentially large volume $(2\pi/4)^{\mathcal{O}(n)} \approx 1.57^{\mathcal{O}(n)}$. This is formally stated below.

Proposition 4 (Exponentially many strictly suboptimal local minima)**.** *Consider*

$$S \subseteq \{0, 1, 2, \ldots, \lfloor n/4 \rfloor - 1\} \tag{385}$$

with $|S| = \Theta(n)$*. For the cost function* $C_S(\vec{\theta})$ *in Eq. (384), there are exponentially many strictly suboptimal local minima* $\{\vec{\theta}_x\}_{x=0}^{2^{|S|}-2}$*, i.e.,*

$$C_S(\vec{\theta}_x) \geq 1 + \min_{\vec{\theta}} C_S(\vec{\theta}), \qquad \textit{(strictly suboptimal)} \tag{386}$$

$$C_S(\vec{\theta}_x) \leq C_S(\vec{\theta}), \quad \forall \left\| \vec{\theta} - \vec{\theta}_x \right\|_\infty < \pi/4, \qquad \textit{(local minimum)} \tag{387}$$

for all $x = 0, \ldots, 2^{|S|} - 2$*.*

Proof. Without loss of generality, we consider n to be divisible by 4. If n is not divisible by 4, we neglect the last $n \bmod 4$ qubits. For convenience, we group and name the parameters $\vec{\theta}$ as follows.

$$\vec{\theta}_{B,j} := (\theta_{1,2j+1}, \theta_{1,2j+2}, \theta_{2,2j+1}, \theta_{3,2j+1}, \theta_{3,2j+2}), \quad \forall j = 0, \ldots, (n/4) - 1, \tag{388}$$

$$\theta_{L,j} := \theta_{2,2j+2}, \quad \forall j = 0, \ldots, (n/4) - 2. \tag{389}$$

Here, $\vec{\theta}_{B,j}$ corresponds to a block of 5 gates acting on 4 qubits. And, $\theta_{L,j}$ corresponds to a single gate linking two blocks. Each integer $x \in \{0, \ldots, 2^{|S|} - 1\}$ corresponds to a local minimum $\vec{\theta}_x$. Let $b_0(x), \ldots, b_{|S|-1}(x)$ be the binary representation of the integer x using $|S|$ bits. We sort the set S from small to large and consider a mapping id from $j \in S$ to the index in S, which is between 0 to $|S| - 1$. The local minimum $\vec{\theta}_x$ is defined as follows. For each $j = 0, \ldots, (n/4) - 1$,

$$\vec{\theta}_{x,B,j} := (\pi/2) \times \begin{cases} (1,1,1,1,1) & \text{if } j \in S \text{ and } b_{\mathrm{id}(j)}(x) = 1 \\ (0,0,0,0,0) & \text{else} \end{cases} \tag{390}$$

And for all $j = 0, \ldots, (n/4) - 2$, $\theta_{x,L,j} := 0$. It is not hard to verify that

$$C_S(\vec{\theta}_x) = 0, \qquad \text{for } x = 2^{|S|} - 1, \tag{391}$$

$$C_S(\vec{\theta}_x) = n - \big(b_0(x) + \ldots + b_{|S|-1}(x)\big) \geq 1, \qquad \text{for } x = 0, \ldots 2^{|S|} - 2. \tag{392}$$

Hence, $\vec{\theta}_{2^{|S|}-1}$ is the global minimum. And for all $x = 0, \ldots, 2^{|S|} - 2$, $\vec{\theta}_x$ is suboptimal. This establishes the first statement of this proposition.

We are now ready to prove the statement that $\vec{\theta}_x$ is a local minimum for all $x = 0, \ldots, 2^{|S|} - 2$. Consider $\vec{\theta}$ such that $\left\| \vec{\theta} - \vec{\theta}_x \right\|_\infty < \pi/4$. We now consider the cost function for each four-qubit block. For block $j \in \{0, \ldots, (n/4) - 1\}$, we have a block of qubits

$$a := 4j + 1, b := 4j + 2, c := 4j + 3, d := 4j + 4. \tag{393}$$

The associated cost function is

$$C_{S,j}(\vec{\theta}) := \mathop{\mathbb{E}}_{|\psi\rangle = \bigotimes_{i=1}^n |\psi_i\rangle \in \mathrm{stab}_1^{\otimes n}} \sum_{i \in \{a,b,c,d\}} \left(1 - \mathrm{Tr}\left(\langle\psi_i| U(\vec{\theta})^\dagger U_S |\psi\rangle\langle\psi| U_S^\dagger U(\vec{\theta}) |\psi_i\rangle\right)\right) \geq 0. \tag{394}$$

If $j \notin S$, or $j \in S$ and $b_{\mathrm{id}(j)}(x) = 1$, we have

$$C_{S,j}(\vec{\theta}_x) = 0 \leq C_{S,j}(\vec{\theta}). \tag{395}$$

So we only need to consider the case when $j \in S$ and $b_{\mathrm{id}(j)}(x) = 0$, which is the case when $U(\vec{\theta}_x)$ acts as identity on block j and U_S acts as a SWAP gate between the first and fourth qubits in block j. In this case, we have the following cost function at $\vec{\theta}_x$,

$$C_{S,j}(\vec{\theta}_x) = 1. \tag{396}$$

For each qubit i, we have the following identity,

$$\mathop{\mathbb{E}}_{|\psi\rangle=\bigotimes_{i=1}^{n}|\psi_i\rangle\in\mathrm{stab}_1^{\otimes n}} \left(1 - \mathrm{Tr}\left(\langle\psi_i| U(\vec{\theta})^\dagger U_S |\psi\rangle\langle\psi| U_S^\dagger U(\vec{\theta}) |\psi_i\rangle\right)\right) \tag{397}$$

$$= \frac{2}{3}\left(1 - \frac{1}{4}\,\mathrm{Tr}_{\neq i}\left(\mathrm{Tr}_i\left(U(\vec{\theta})U_S\right)^\dagger \left(\frac{I_{n-1}}{2^{n-1}}\right) \mathrm{Tr}_i\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right)\right), \tag{398}$$

where $\frac{I_{n-1}}{2^{n-1}}$ is the maximally mixed state over $n-1$ qubits. By the definition of U_S and $U(\vec{\theta})$, $U(\vec{\theta})^\dagger U_S$ is a linear combination of permutation operators with complex-valued weights. For $i = a$, we can rewrite the tensor contractions in Eq. (398) using the three gates associated with parameters $\theta_{B,j,2}, \theta_{B,j,3}, \theta_{B,j,4}$. By first treating the maximally mixed states and the tracing operation $\mathrm{Tr}_{\neq i}$, we can rewrite the three gates as depolarizing channels, which gives rise to the following identity.

$$\frac{1}{4}\,\mathrm{Tr}_{\neq a}\left(\mathrm{Tr}_a\left(U(\vec{\theta})U_S\right)^\dagger \left(\frac{I_{n-1}}{2^{n-1}}\right) \mathrm{Tr}_a\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) = \lambda_a + (1-\lambda_a)\frac{1}{4}, \tag{399}$$

where $\lambda_a := \sin(\theta_{B,j,2})^2 \sin(\theta_{B,j,3})^2 \sin(\theta_{B,j,4})^2$. Similarly, for $i = d$, we have

$$\frac{1}{4}\,\mathrm{Tr}_{\neq d}\left(\mathrm{Tr}_d\left(U(\vec{\theta})U_S\right)^\dagger \left(\frac{I_{n-1}}{2^{n-1}}\right) \mathrm{Tr}_d\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) = \lambda_d + (1-\lambda_d)\frac{1}{4}, \tag{400}$$

where $\lambda_d := \sin(\theta_{B,j,1})^2 \sin(\theta_{B,j,3})^2 \sin(\theta_{B,j,5})^2$. For $i = b$, the tensor contractions in in Eq. (398) using the four gates associated with parameters $\theta_{B,j,1}, \theta_{B,j,3}, \theta_{B,j,4}, \theta_{L,j-1}$. We can rewrite the two gates associated with $\theta_{L,j-1}$ and $\theta_{B,j,3}$ in terms of depolarizing channels on qubit a, b, respectively. By enumerating all possible terms, we have

$$\frac{1}{4}\,\mathrm{Tr}_{\neq b}\left(\mathrm{Tr}_b\left(U(\vec{\theta})U_S\right)^\dagger \left(\frac{I_{n-1}}{2^{n-1}}\right) \mathrm{Tr}_b\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) \tag{401}$$

$$= \cos(\theta_{B,j,1})^2 \cos(\theta_{B,j,4})^2 \left(\cos(\theta_{B,j,3})^2 + \frac{1}{4}\sin(\theta_{B,j,3})^2\right) \tag{402}$$

$$+ \sin(\theta_{B,j,1})^2 \sin(\theta_{B,j,4})^2 \left(\cos(\theta_{L,j-1})^2 + \frac{1}{4}\sin(\theta_{L,j-1})^2\right) \tag{403}$$

$$+ \frac{1}{4}\left(\cos(\theta_{B,j,1})^2 \sin(\theta_{B,j,4})^2 + \sin(\theta_{B,j,1})^2 \cos(\theta_{B,j,4})^2\right) \tag{404}$$

$$- \frac{3}{2}\cos(\theta_{B,j,1})\sin(\theta_{B,j,1})\cos(\theta_{B,j,4})\sin(\theta_{B,j,4})\cos(\theta_{L,j-1})^2\cos(\theta_{B,j,3})^2 \tag{405}$$

$$\leq \cos(\theta_{B,j,1})^2 \cos(\theta_{B,j,4})^2 \left(1 - \frac{3}{4}\sin(\theta_{B,j,3})^2\right) + \sin(\theta_{B,j,1})^2 \sin(\theta_{B,j,4})^2 \tag{406}$$

$$+ \frac{1}{4}\left(\cos(\theta_{B,j,1})^2 \sin(\theta_{B,j,4})^2 + \sin(\theta_{B,j,1})^2 \cos(\theta_{B,j,4})^2\right) \tag{407}$$

$$+ \frac{3}{2}\left|\cos(\theta_{B,j,1})\sin(\theta_{B,j,1})\cos(\theta_{B,j,4})\sin(\theta_{B,j,4})\right|\left(1 - \sin(\theta_{B,j,3})^2\right). \tag{408}$$

Because $\left\|\vec{\theta}-\vec{\theta}_x\right\|_\infty < \pi/4$, we have $\cos(\theta_{B,j,1}) \geq 0, \cos(\theta_{B,j,4}) \geq 0$ and

$$|\sin(\theta_{B,j,1})| = \sin(|\theta_{B,j,1}|), \quad |\sin(\theta_{B,j,4})| = \sin(|\theta_{B,j,4}|). \tag{409}$$

We can use trigonometric identities to obtain

$$\frac{1}{4}\operatorname{Tr}_{\neq b}\left(\operatorname{Tr}_b\left(U(\vec{\theta})U_S\right)^\dagger\left(\frac{I_{n-1}}{2^{n-1}}\right)\operatorname{Tr}_b\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) \tag{410}$$

$$\leq 1-\frac{3}{4}\sin\left(|\theta_{B,j,1}|-|\theta_{B,j,4}|\right)^2-\frac{3}{4}\cos(\theta_{B,j,1})^2\cos(\theta_{B,j,4})^2\sin(\theta_{B,j,3})^2 \tag{411}$$

$$-\frac{3}{2}\left|\cos(\theta_{B,j,1})\sin(\theta_{B,j,1})\cos(\theta_{B,j,4})\sin(\theta_{B,j,4})\right|\sin(\theta_{B,j,3})^2 \tag{412}$$

$$\leq 1-\frac{3}{4}\sin(\theta_{B,j,3})^2\cos(\theta_{B,j,1})^2\cos(\theta_{B,j,4})^2. \tag{413}$$

Similarly, we have

$$\frac{1}{4}\operatorname{Tr}_{\neq c}\left(\operatorname{Tr}_c\left(U(\vec{\theta})U_S\right)^\dagger\left(\frac{I_{n-1}}{2^{n-1}}\right)\operatorname{Tr}_c\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) \tag{414}$$

$$\leq 1-\frac{3}{4}\sin(\theta_{B,j,3})^2\cos(\theta_{B,j,2})^2\cos(\theta_{B,j,5})^2. \tag{415}$$

Combining all four upper bounds on

$$\frac{1}{4}\operatorname{Tr}_{\neq i}\left(\operatorname{Tr}_i\left(U(\vec{\theta})U_S\right)^\dagger\left(\frac{I_{n-1}}{2^{n-1}}\right)\operatorname{Tr}_i\left(U(\vec{\theta})^\dagger U_S\right)^\dagger\right) \tag{416}$$

for $i = a, b, c, d$, we can obtain the cost function associated to this block,

$$C_{S,j}(\vec{\theta}) \geq 1-\frac{1}{2}\sin(\theta_{B,j,2})^2\sin(\theta_{B,j,3})^2\sin(\theta_{B,j,4})^2-\frac{1}{2}\sin(\theta_{B,j,1})^2\sin(\theta_{B,j,3})^2\sin(\theta_{B,j,5})^2 \tag{417}$$

$$+\frac{1}{2}\sin(\theta_{B,j,3})^2\cos(\theta_{B,j,1})^2\cos(\theta_{B,j,4})^2+\frac{1}{2}\sin(\theta_{B,j,3})^2\cos(\theta_{B,j,2})^2\cos(\theta_{B,j,5})^2. \tag{418}$$

From $\left\|\vec{\theta}-\vec{\theta}_x\right\|_\infty < \pi/4$, we have

$$|\sin(\theta_{B,j,k})| < 0.5, \quad \forall k = 1,2,3,4,5, \tag{419}$$

$$|\cos(\theta_{B,j,k})| > 0.5, \quad \forall k = 1,2,3,4,5. \tag{420}$$

Hence, $C_{S,j}(\vec{\theta}) \geq 1 = C_{S,j}(\vec{\theta}_x)$. Together with the fact that

$$C_S(\vec{\theta}) = \sum_{j=0}^{(n/4)-1} C_{S,j}(\vec{\theta}), \tag{421}$$

we have established the claim $C_S(\vec{\theta}_x) \leq C_S(\vec{\theta})$. □

References

1. Bravyi, S., Gosset, D. & Koenig, R. Quantum advantage with shallow circuits. *Science* **362,** 308–311 (2018).

2. Bravyi, S., Gosset, D., Koenig, R. & Tomamichel, M. Quantum advantage with noisy shallow circuits. *Nature Physics* **16,** 1040–1045 (2020).

3. Watts, A. B. & Parham, N. Unconditional Quantum Advantage for Sampling with Shallow Circuits. *arXiv preprint arXiv:2301.00995* (2023).

4. Watts, A. B., Kothari, R., Schaeffer, L. & Tal, A. *Exponential separation between shallow quantum circuits and unbounded fan-in shallow classical circuits* in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019), 515–526.

5. Terhal, B. M. & DiVincenzo, D. P. *Adaptive Quantum Computation, Constant Depth Quantum Circuits and Arthur-Merlin Games* 2004. arXiv: quant-ph/0205133 [quant-ph].

6. Gao, X., Wang, S.-T. & Duan, L.-M. Quantum Supremacy for Simulating a Translation-Invariant Ising Spin Model. *Phys. Rev. Lett.* **118,** 040502 (4 Jan. 2017).

7. Bermejo-Vega, J., Hangleiter, D., Schwarz, M., Raussendorf, R. & Eisert, J. Architectures for quantum simulation showing a quantum speedup. *Physical Review X* **8,** 021010 (2018).

8. Haferkamp, J., Hangleiter, D., Bouland, A., Fefferman, B., Eisert, J. & Bermejo-Vega, J. Closing gaps of a quantum advantage with short-time hamiltonian dynamics. *Physical Review Letters* **125,** 250501 (2020).

9. Hangleiter, D. & Eisert, J. Computational advantage of quantum random sampling. *Reviews of Modern Physics* **95,** 035001 (2023).

10. Farhi, E. & Neven, H. Classification with quantum neural networks on near term processors. *arXiv preprint arXiv:1802.06002* (2018).

11. Benedetti, M., Lloyd, E., Sack, S. & Fiorentini, M. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology* **4,** 043001 (2019).

12. Beer, K., Bondarenko, D., Farrelly, T., Osborne, T. J., Salzmann, R., Scheiermann, D. & Wolf, R. Training deep quantum neural networks. *Nature communications* **11,** 808 (2020).

13. Bausch, J. Recurrent quantum neural networks. *Advances in neural information processing systems* **33,** 1368–1379 (2020).

14. Skolik, A., McClean, J. R., Mohseni, M., van der Smagt, P. & Leib, M. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence* **3,** 1–11 (2021).

15. Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A. & Woerner, S. The power of quantum neural networks. *Nature Computational Science* **1,** 403–409 (2021).

16. Caro, M. C., Huang, H.-Y., Cerezo, M., Sharma, K., Sornborger, A., Cincio, L. & Coles, P. J. Generalization in quantum machine learning from few training data. *Nature communications* **13,** 4919 (2022).

17. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications* **12,** 1791 (2021).

18. Ostaszewski, M., Grant, E. & Benedetti, M. Structure optimization for parameterized quantum circuits. *Quantum* **5,** 391 (2021).

19. Pesah, A., Cerezo, M., Wang, S., Volkoff, T., Sornborger, A. T. & Coles, P. J. Absence of Barren Plateaus in Quantum Convolutional Neural Networks. *Phys. Rev. X* **11,** 041011 (4 Oct. 2021).

20. Du, Y., Hsieh, M.-H., Liu, T., You, S. & Tao, D. Learnability of quantum neural networks. *PRX Quantum* **2,** 040337 (2021).

21. Holmes, Z., Sharma, K., Cerezo, M. & Coles, P. J. Connecting ansatz expressibility to gradient magnitudes and barren plateaus. *PRX Quantum* **3,** 010313 (2022).

22. Sharma, K., Cerezo, M., Cincio, L. & Coles, P. J. Trainability of dissipative perceptron-based quantum neural networks. *Physical Review Letters* **128,** 180505 (2022).

23. Anschuetz, E. R. & Kiani, B. T. Quantum variational algorithms are swamped with traps. *Nature Communications* **13,** 7760 (2022).

24. Cerezo, M., Verdon, G., Huang, H.-Y., Cincio, L. & Coles, P. J. Challenges and opportunities in quantum machine learning. *Nature Computational Science* **2,** 567–576 (2022).

25. Linial, N., Mansour, Y. & Nisan, N. Constant depth circuits, Fourier transform, and learnability. *Journal of the ACM (JACM)* **40,** 607–620 (1993).

26. Mossel, E., O'Donnell, R. & Servedio, R. P. *Learning Juntas* in *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, San Diego, CA, USA, 2003), 206–212.

27. Carmosino, M. L., Impagliazzo, R., Kabanets, V. & Kolokolova, A. *Learning Algorithms from Natural Proofs* in *31st Conference on Computational Complexity (CCC 2016)* (ed Raz, R.) **50** (Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2016), 10:1–10:24.

28. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature communications* **9,** 4812 (2018).

29. Holmes, Z., Arrasmith, A., Yan, B., Coles, P. J., Albrecht, A. & Sornborger, A. T. Barren plateaus preclude learning scramblers. *Physical Review Letters* **126,** 190501 (2021).

30. Wang, S., Fontana, E., Cerezo, M., Sharma, K., Sone, A., Cincio, L. & Coles, P. J. Noise-induced barren plateaus in variational quantum algorithms. *Nature communications* **12,** 6961 (2021).

31. Chen, S., Cotler, J., Huang, H.-Y. & Li, J. The complexity of NISQ. *arXiv preprint arXiv:2210.07234* (2022).

32. Cincio, L., Subaşı, Y., Sornborger, A. T. & Coles, P. J. Learning the quantum algorithm for state overlap. *New Journal of Physics* **20,** 113022 (2018).

33. Khatri, S., LaRose, R., Poremba, A., Cincio, L., Sornborger, A. T. & Coles, P. J. Quantum-assisted quantum compiling. *Quantum* **3,** 140 (2019).

34. Sharma, K., Khatri, S., Cerezo, M. & Coles, P. J. Noise resilience of variational quantum compiling. *New Journal of Physics* **22,** 043006 (2020).

35. Jones, T. & Benjamin, S. C. Robust quantum compilation and circuit optimisation via energy minimisation. *Quantum* **6,** 628 (2022).

36. Cirstoiu, C., Holmes, Z., Iosue, J., Cincio, L., Coles, P. J. & Sornborger, A. Variational fast forwarding for quantum simulation beyond the coherence time. *npj Quantum Information* **6,** 82 (2020).

37. Yao, Y.-X., Gomes, N., Zhang, F., Wang, C.-Z., Ho, K.-M., Iadecola, T. & Orth, P. P. Adaptive variational quantum dynamics simulations. *PRX Quantum* **2,** 030307 (2021).

38. Gibbs, J., Holmes, Z., Caro, M. C., Ezzell, N., Huang, H.-Y., Cincio, L., Sornborger, A. T. & Coles, P. J. Dynamical simulation via quantum machine learning with provable generalization. *arXiv preprint arXiv:2204.10269* (2022).

39. Caro, M. C., Huang, H.-Y., Ezzell, N., Gibbs, J., Sornborger, A. T., Cincio, L., Coles, P. J. & Holmes, Z. Out-of-distribution generalization for learning quantum dynamics. *Nature Communications* **14,** 3751 (2023).

40. Jerbi, S., Gibbs, J., Rudolph, M. S., Caro, M. C., Coles, P. J., Huang, H.-Y. & Holmes, Z. The power and limitations of learning quantum dynamics incoherently. *arXiv preprint arXiv:2303.12834* (2023).

41. Lloyd, S. & Weedbrook, C. Quantum generative adversarial learning. *Physical review letters* **121,** 040502 (2018).

42. Benedetti, M., Garcia-Pintos, D., Perdomo, O., Leyton-Ortega, V., Nam, Y. & Perdomo-Ortiz, A. A generative modeling approach for benchmarking and training shallow quantum circuits. *npj Quantum Information* **5,** 45 (2019).

43. Coyle, B., Mills, D., Danos, V. & Kashefi, E. The Born supremacy: quantum advantage and training of an Ising Born machine. *npj Quantum Information* **6,** 60 (2020).

44. Gao, X., Anschuetz, E. R., Wang, S.-T., Cirac, J. I. & Lukin, M. D. Enhancing generative models via quantum correlations. *Physical Review X* **12,** 021037 (2022).

45. Rudolph, M. S., Toussaint, N. B., Katabarwa, A., Johri, S., Peropadre, B. & Perdomo-Ortiz, A. Generation of high-resolution handwritten digits with an ion-trap quantum computer. *Physical Review X* **12,** 031010 (2022).

46. Zhu, E. Y., Johri, S., Bacon, D., Esencan, M., Kim, J., Muir, M., Murgai, N., Nguyen, J., Pisenti, N., Schouela, A., *et al.* Generative quantum learning of joint probability distribution functions. *Physical Review Research* **4,** 043092 (2022).

47. Cramer, M., Plenio, M. B., Flammia, S. T., Somma, R., Gross, D., Bartlett, S. D., Landon-Cardinal, O., Poulin, D. & Liu, Y.-K. Efficient quantum state tomography. *Nature communications* **1,** 149 (2010).

48. Lanyon, B., Maier, C., Holzäpfel, M., Baumgratz, T., Hempel, C., Jurcevic, P., Dhand, I., Buyskikh, A., Daley, A., Cramer, M., *et al.* Efficient tomography of a quantum many-body system. *Nature Physics* **13,** 1158–1162 (2017).

49. Gebhart, V., Santagati, R., Gentile, A. A., Gauger, E. M., Craig, D., Ares, N., Banchi, L., Marquardt, F., Pezzè, L. & Bonato, C. Learning quantum systems. *Nature Reviews Physics* **5,** 141–156 (2023).

50. Anshu, A., Arunachalam, S., Kuwahara, T. & Soleimanifar, M. *Sample-efficient learning of quantum many-body systems* in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)* (2020), 685–691.

51. Rouzé, C. & França, D. S. Learning quantum many-body systems from a few copies. *arXiv preprint arXiv:2107.03333* (2021).

52. Haah, J., Kothari, R. & Tang, E. *Optimal learning of quantum Hamiltonians from high-temperature Gibbs states* in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (2022), 135–146.

53. Montanaro, A. Learning stabilizer states by Bell sampling. *arXiv preprint arXiv:1707.04012* (2017).

54. Gross, D., Nezami, S. & Walter, M. Schur–Weyl duality for the Clifford group with applications: Property testing, a robust Hudson theorem, and de Finetti representations. *Communications in Mathematical Physics* **385,** 1325–1393 (2021).

55. Grewal, S., Iyer, V., Kretschmer, W. & Liang, D. Low-Stabilizer-Complexity Quantum States Are Not Pseudorandom. *arXiv preprint arXiv:2209.14530* (2022).

56. Grewal, S., Iyer, V., Kretschmer, W. & Liang, D. Improved Stabilizer Estimation via Bell Difference Sampling. *arXiv preprint arXiv:2304.13915* (2023).

57. Arunachalam, S., Bravyi, S., Dutt, A. & Yoder, T. J. Optimal algorithms for learning quantum phase states. *arXiv preprint arXiv:2208.07851* (2022).

58. Aaronson, S. & Grewal, S. *Efficient Tomography of Non-Interacting-Fermion States* in *18th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2023)* (2023).

59. Lai, C.-Y. & Cheng, H.-C. Learning quantum circuits of some T gates. *IEEE Transactions on Information Theory* **68,** 3951–3964 (2022).

60. Flammia, S. T. & Wallman, J. J. Efficient estimation of Pauli channels. *ACM Transactions on Quantum Computing* **1,** 1–32 (2020).

61. Flammia, S. T. & O'Donnell, R. Pauli error estimation via population recovery. *Quantum* **5,** 549 (2021).

62. Chen, S., Zhou, S., Seif, A. & Jiang, L. Quantum advantages for Pauli channel estimation. *Physical Review A* **105,** 032435 (2022).

63. Van Den Berg, E., Minev, Z. K., Kandala, A. & Temme, K. Probabilistic error cancellation with sparse Pauli–Lindblad models on noisy quantum processors. *Nature Physics,* 1–6 (2023).

64. Li, Z., Zou, L. & Hsieh, T. H. Hamiltonian tomography via quantum quench. *Physical review letters* **124,** 160502 (2020).

65. Che, L., Wei, C., Huang, Y., Zhao, D., Xue, S., Nie, X., Li, J., Lu, D. & Xin, T. Learning quantum Hamiltonians from single-qubit measurements. *Physical Review Research* **3,** 023246 (2021).

66. Yu, W., Sun, J., Han, Z. & Yuan, X. *Practical and Efficient Hamiltonian Learning* 2022. arXiv: 2201.00190 [quant-ph].

67. Hangleiter, D., Roth, I., Eisert, J. & Roushan, P. *Precise Hamiltonian identification of a superconducting quantum processor* 2021. arXiv: 2108.08319 [quant-ph].

68. Franca, D. S., Markovich, L. A., Dobrovitski, V., Werner, A. H. & Borregaard, J. Efficient and robust estimation of many-qubit Hamiltonians. *arXiv preprint arXiv:2205.09567* (2022).

69. Zubida, A., Yitzhaki, E., Lindner, N. H. & Bairey, E. Optimal short-time measurements for Hamiltonian learning. *arXiv preprint arXiv:2108.08824* (2021).

70. Bairey, E., Arad, I. & Lindner, N. H. Learning a local Hamiltonian from local measurements. *Physical review letters* **122,** 020504 (2019).

71. Granade, C. E., Ferrie, C., Wiebe, N. & Cory, D. G. Robust online Hamiltonian learning. *New Journal of Physics* **14,** 103013 (2012).

72. Gu, A., Cincio, L. & Coles, P. J. Practical Black Box Hamiltonian Learning. *arXiv preprint arXiv:2206.15464* (2022).

73. Wilde, F., Kshetrimayum, A., Roth, I., Hangleiter, D., Sweke, R. & Eisert, J. *Scalably learning quantum many-body Hamiltonians from dynamical data* 2022.

74. Huang, H.-Y., Tong, Y., Fang, D. & Su, Y. Learning many-body Hamiltonians with Heisenberg-limited scaling. *Physical Review Letters* **130,** 200403 (2023).

75. Anshu, A. & Arunachalam, S. *A survey on the complexity of learning quantum states* 2023. arXiv: 2305.20069 [quant-ph].

76. Terhal, B. M. & DiVincenzo, D. P. Classical simulation of noninteracting-fermion quantum circuits. *Physical Review A* **65,** 032325 (2002).

77. Aaronson, S. & Gottesman, D. Improved simulation of stabilizer circuits. *Physical Review A* **70,** 052328 (2004).

78. Cirac, J. I., Perez-Garcia, D., Schuch, N. & Verstraete, F. Matrix product states and projected entangled pair states: Concepts, symmetries, theorems. *Reviews of Modern Physics* **93,** 045003 (2021).

79. Wild, D. S. & Alhambra, Á. M. Classical simulation of short-time quantum dynamics. *PRX Quantum* **4,** 020340 (2023).

80. Yin, C. & Lucas, A. Polynomial-time classical sampling of high-temperature quantum Gibbs states. *arXiv preprint arXiv:2305.18514* (2023).

81. Bermejo-Vega, J., Hangleiter, D., Schwarz, M., Raussendorf, R. & Eisert, J. Architectures for Quantum Simulation Showing a Quantum Speedup. *Phys. Rev. X* **8,** 021010 (2 Apr. 2018).

82. Aaronson, S. *Shadow tomography of quantum states* in *STOC* (2018), 325–338.

83. Bădescu, C. & O'Donnell, R. Improved quantum data analysis. *arXiv preprint arXiv:2011.10908* (2020).

84. Huang, H.-Y., Kueng, R. & Preskill, J. Predicting many properties of a quantum system from very few measurements. *Nature Physics* **16,** 1050–1057 (Oct. 2020).

85. Levy, R., Luo, D. & Clark, B. K. Classical shadows for quantum process tomography on near-term quantum computers. *arXiv preprint arXiv:2110.02965* (2021).

86. Huang, H.-Y., Chen, S. & Preskill, J. Learning to predict arbitrary quantum processes. *arXiv preprint arXiv:2210.14894* (2022).

87. Kunjummen, J., Tran, M. C., Carney, D. & Taylor, J. M. Shadow process tomography of quantum channels. *Physical Review A* **107,** 042403 (2023).

88. Elben, A., Flammia, S. T., Huang, H.-Y., Kueng, R., Preskill, J., Vermersch, B. & Zoller, P. The randomized measurement toolbox. *arXiv preprint arXiv:2203.11374* (2022).

89. Schumacher, B. & Werner, R. F. *Reversible quantum cellular automata* 2004. arXiv: quant-ph/0405174 [quant-ph].

90. Gross, D., Nesme, V., Vogts, H. & Werner, R. F. Index Theory of One Dimensional Quantum Walks and Cellular Automata. *Communications in Mathematical Physics* **310,** 419–454 (Jan. 2012).

91. Haah, J., Fidkowski, L. & Hastings, M. B. Nontrivial Quantum Cellular Automata in Higher Dimensions. *Communications in Mathematical Physics* **398,** 469–540 (Nov. 2022).

92. Shirley, W., Chen, Y.-A., Dua, A., Ellison, T. D., Tantivasadakarn, N. & Williamson, D. J. Three-Dimensional Quantum Cellular Automata from Chiral Semion Surface Topological Order and beyond. *PRX Quantum* **3,** 030326 (3 Aug. 2022).

93. Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S. & Eisert, J. Quantum state tomography via compressed sensing. *Physical review letters* **105,** 150401 (2010).

94. Yu, N. & Wei, T.-C. *Learning marginals suffices!* 2023. arXiv: 2303.08938 [quant-ph].

95. Zalka, C. Grover's quantum searching algorithm is optimal. *Physical Review A* **60,** 2746 (1999).

96. Nielsen, M. A. A simple formula for the average gate fidelity of a quantum dynamical operation. *Physics Letters A* **303,** 249–252 (2002).

97. Montanaro, A. & de Wolf, R. A survey of quantum property testing. *arXiv preprint arXiv:1310.2035* (2013).

98. Duan, R., Feng, Y. & Ying, M. Perfect distinguishability of quantum operations. *Physical Review Letters* **103,** 210501 (2009).

99. Bengtsson, I. & Życzkowski, K. *Geometry of quantum states: an introduction to quantum entanglement* (Cambridge university press, 2017).

100. Haah, J., Kothari, R., O'Donnell, R. & Tang, E. Query-optimal estimation of unitary channels in diamond distance. *arXiv preprint arXiv:2302.14066* (2023).

101. Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D. P., Margolus, N., Shor, P., Sleator, T., Smolin, J. A. & Weinfurter, H. Elementary gates for quantum computation. *Phys. Rev. A* **52,** 3457–3467 (5 Nov. 1995).

102. Shende, V. V., Bullock, S. S. & Markov, I. L. *Synthesis of quantum logic circuits* in *Proceedings of the 2005 Asia and South Pacific Design Automation Conference* (2005), 272–275.

103. Watrous, J. *The theory of quantum information* (Cambridge university press, 2018).

104. Chen, T., Nadimpalli, S. & Yuen, H. *Testing and learning quantum juntas nearly optimally* in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2023), 1163–1185.

105. Brandão, F. G. S. L. & Kastoryano, M. J. Finite Correlation Length Implies Efficient Preparation of Quantum Thermal States. *Communications in Mathematical Physics* **365,** 1–16 (Jan. 2019).

106. Bennett, C. H., Bernstein, E., Brassard, G. & Vazirani, U. Strengths and weaknesses of quantum computing. *SIAM journal on Computing* **26,** 1510–1523 (1997).

107. Lieb, E. H. & Ruskai, M. B. Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics* **14,** 1938–1941. eprint: https://pubs.aip.org/aip/jmp/article-pdf/14/12/1938/8146113/1938_1_online.pdf (Nov. 2003).

108. Surawy-Stepney, T., Kahn, J., Kueng, R. & Guta, M. Projected least-squares quantum process tomography. *Quantum* **6,** 844 (2022).

109. Caro, M. C., Huang, H.-Y., Ezzell, N., Gibbs, J., Sornborger, A. T., Cincio, L., Coles, P. J. & Holmes, Z. Out-of-distribution generalization for learning quantum dynamics. *Nature Communications* **14,** 3751 (2023).