Foundations for learning from noisy quantum experiments

Hsin-Yuan Huang,[1,2] Steven T. Flammia,[2] and John Preskill[1,2]

[1]*Institute for Quantum Information and Matter and Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA, USA*
[2]*AWS Center for Quantum Computing, Pasadena, CA, USA*

(Dated: February 28, 2022)

We study our ability to learn physical operations in quantum systems where all operations, from state preparation, dynamics, to measurement, are *a priori* unknown. We prove that without any prior knowledge, if one can explore the full quantum state space by composing the operations, then every operation could be learned up to an arbitrarily small error. When one cannot explore the full space but the operations are approximately known, we present an efficient algorithm for learning all operations up to a single unlearnable parameter corresponding to the fidelity of the initial state assuming gate-independent noise on Clifford gates. Our algorithm for learning the Clifford gate noise uses a number of experiments linear in the number of parameters, which is quadratically fewer than the best known randomized benchmarking protocol. When these assumptions are not met, the true description of the noise can be fundamentally unlearnable, e.g., we prove that no benchmarking protocol can learn the Pauli noise on Clifford+T gates if the Pauli noise depends on the gates. Even if the noise cannot be learned, we prove that a large quantum advantage can be achieved in a recent learning task performed on the Sycamore quantum processor.

Understanding what we can learn from experiments is central to many scientific fields. By conducting experiments we obtain information about the physical world. This information can be organized into knowledge allowing us to predict how the world would behave under different circumstances and to design complex systems with desired functionalities. To develop quantum technology, understanding what we can learn from quantum experiments is crucial.

In this work, we build on computational learning theory [19, 32, 37] to develop a rigorous theory for reasoning about what can and cannot be learned. We represent scientists and their classical algorithms abstractly as classical agents that conduct experiments by specifying actions that control a quantum mechanical system. Actions include specifying which initial states are prepared, which CPTP maps are performed, and which POVM measurements are executed at the end of an experiment. Classical agents do not have perfect knowledge about how these actions affect the quantum world, e.g., what state is prepared, or what quantum processes and measurements are actually implemented. But classical agents can improve their understanding of the physical world through experiments. We refer to the mapping from each action to the corresponding physical operation as a world model $\mathcal{W}$. The central question we would like to answer is: What can a classical agent learn about the true world model $\mathcal{W}^*$ describing the underlying physical reality?

To answer this question, we make a distinction between two different viewpoints for assessing how well the classical agent learns. The first viewpoint judges whether the classical agent accurately *learns the intrinsic descriptions* of the physical operations. This is the viewpoint commonly considered in quantum tomography [2, 3, 12, 21–23, 27, 29, 38]. For example, given a quantum computer, we might want the classical agent to characterize the noise afflicting the state preparations, maps, or measurements acting on various qubits. Knowledge about the intrinsic descriptions of the physical operations is crucial for calibrating, controlling, and improving a complex quantum many-body system [8, 50, 51]. If, for example, the classical agent finds the measurements to be particularly noisy, we should focus on improving our measurement procedure.

The second viewpoint examines whether the classical agent can *predict the extrinsic behavior* of the quantum system. In particular, given an experiment, we ask the classical agent to predict the experimental outcomes. The classical agent does not need to learn the actual descriptions of the physical operations and is free to use any model as long as the prediction is accurate. The second viewpoint is relevant when we want to control the quantum system to achieve specific tasks, such as mitigating errors in a particular computation [14, 31, 44, 46]. Because the true physical descriptions are

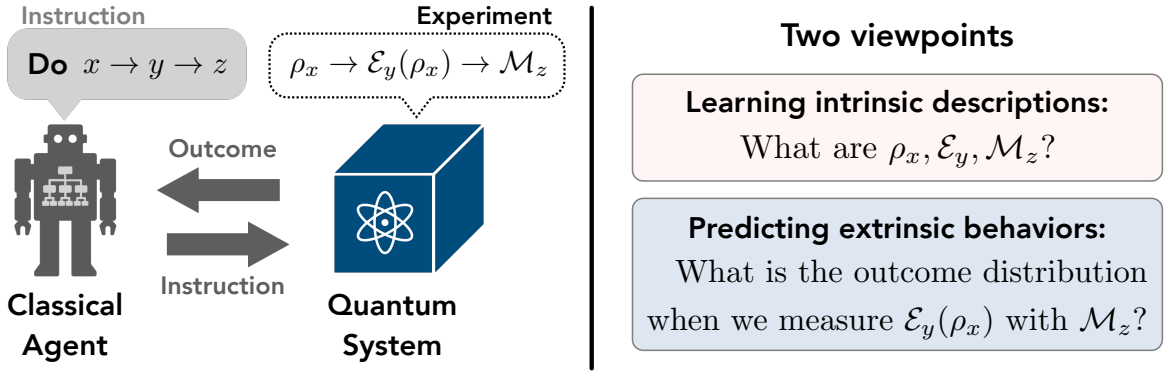

Figure 1: Illustration of a classical agent that can learn from quantum experiments. Classical agent specify an instruction consisting of the actions x, y, z for the experiment. The quantum system runs the experiment using a state ρ_x, a quantum process $\mathcal{E}_y$, and a POVM measurement $\mathcal{M}_z$ to produce an experimental outcome. The classical agent does not have prior knowledge on what $\rho_x, \mathcal{E}_y, \mathcal{M}_z$ are. The goal of the classical agent is to either learn $\rho_x, \mathcal{E}_y, \mathcal{M}_z$ (learning intrinsic descriptions) or predict the distribution of the experimental outcome (predicting extrinsic behaviors).

not learned, a model that can predict extrinsic behavior might not provide useful guidance regarding how to improve the quantum device. While intrinsic descriptions are more informative than a model for predicting extrinsic behavior, intrinsic descriptions are also much more challenging to learn.

In this work, we will provide fundamental results for understanding what one could learn from noisy quantum experiments. We will present case studies that illustrate the practical implications of the foundations established in this work. In a subsequent work, we will present a versatile mathematical framework for developing rigorous neural network algorithms capable of learning a wide range of noise models in quantum many-body systems.

I. Foundations

To elucidate the basic features of learning intrinsic descriptions and predicting extrinsic behaviors, we prove a series of fundamental results regarding what the classical agent can learn and how efficiently the classical agent can learn it. We begin by establishing the formalism. Then, we present the fundamental results that elucidate the two viewpoints: learning intrinsic descriptions and predicting extrinsic behaviors.

A. A formalism for learning from quantum experiments

We consider three sets of actions that the classical agent could play: $\mathcal{X}$ is the set of actions for preparing different kind of states, $\mathcal{Y}$ consists of actions for implementing certain physical evolutions, and $\mathcal{Z}$ is a set of possible POVM measurements. A world model $\mathcal{W}$ is a mapping from the actions to the actual physical operations. For example, each action $y \in \mathcal{Y}$ corresponds to a quantum channel $\mathcal{E}_y$ that will be applied to the quantum system when the classical agent performs the action y. We represent the world model by the collection of all the physical operations associated to every action,

$$\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}}), \tag{1}$$

where ρ_x is a quantum state, $\mathcal{E}_y$ is a quantum channel, and $\mathcal{M}_z$ is a POVM measurement. In Appendix A, we provide formal definitions of world models and the relations between world models. Due to the intrinsic degeneracy in the mathematical formulation of quantum mechanics, two world models related by a unitary or anti-unitary transformation describe the same physical reality.

The classical agent conduct experiments by controlling the quantum system with the given set of actions. Each experiment E begins with an action $x \in \mathcal{X}$ that prepares the state ρ_x, followed by

a sequence of actions $y_1, \ldots, y_L \in \mathcal{Y}$ that evolves the state ρ_x to $\mathcal{E}_{y_L}(\ldots \mathcal{E}_{y_2}(\mathcal{E}_{y_1}(\rho_x))\ldots)$, and the experiment ends with an action $z \in \mathcal{Z}$ that produces a measurement outcome after performing the POVM $\mathcal{M}_z$ on $\mathcal{E}_{y_L}(\ldots \mathcal{E}_{y_2}(\mathcal{E}_{y_1}(\rho_x))\ldots)$. Suppose the true world model is $\mathcal{W}_{\text{true}}$, and suppose the classical agent has a class of candidate models $\mathcal{Q} = \{\mathcal{W}\}$ for the true world model $\mathcal{W}_{\text{true}} \in \mathcal{Q}$ encoding his/her prior knowledge. The goal of the classical agent is to obtain knowledge (learn) about the true world model $\mathcal{W}_{\text{true}}$ in the model class $\mathcal{Q}$ by conducting experiments. The set of candidate models $\mathcal{Q}$ could be *uncountably large*, which is common in actual experimental setup as the possible noise process forms an uncountably infinite set. The set of models $\mathcal{Q}$ will be referred to as the *model class*, following the nomenclature in classical learning theory [37].

B. Learning intrinsic descriptions of the world model

A common viewpoint for judging the success of the classical learning agent is to see if he/she could identify the physical description of the true world model $\mathcal{W}_{\text{true}} \in \mathcal{Q}$ (up to a unitary or anti-unitary transformation). In particular, given any error $\epsilon > 0$, we check whether the classical agent can learn to describe the states ρ_x, the maps $\mathcal{E}_y$, and the POVMs $\mathcal{M}_z$ in $\mathcal{W}_{\text{true}}$ to error at most ϵ. Formal definition of learnability is given in Appendix C. Such a viewpoint is often taken in the literature of quantum tomography to provide rigorous performance guarantee. In quantum state/process tomography, we assume that certain actions are perfectly known and we would like to learn the descriptions of the states or the CPTP maps to any error. Here, all actions may be unknown, but we would still like to learn the true physical descriptions to arbitrarily small error.

Learning the intrinsic physical description is challenging because the classical agent only observe the behavior of the world model under the given set of actions. If the actions are limited, then he/she is unable to learn the intrinsic physical description of $\mathcal{W}_{\text{true}}$. For example, a classical agent cannot distinguish between the following two distinct physical realities in a single-qubit system,

$$\mathcal{W}^A: \quad \rho_0^A = I/2, \qquad \mathcal{E}_H^A(\rho) = H\rho H^\dagger, \qquad \mathcal{E}_T^A(\rho) = T\rho T^\dagger, \qquad \mathcal{M}_0^A = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{2}$$

$$\mathcal{W}^B: \quad \rho_0^B = I/2, \qquad \mathcal{E}_H^B(\rho) = I/2, \qquad \mathcal{E}_T^B(\rho) = I/2, \qquad \mathcal{M}_0^B = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{3}$$

with actions $\mathcal{X} = \{0\}, \mathcal{Y} = \{H, T\}, \mathcal{Z} = \{0\}$. In both world A and B, the experimental outcome is always a uniform distribution over the two outcomes $0, 1$. Hence, given any *possibly uncountably large* model class $\mathcal{Q}$ that contains $\mathcal{W}^A$ and $\mathcal{W}^B$, the classical agent would not be able to learn all actions to arbitrarily small error. This stems from the fact that the actions are too limited to tell the two physical realities apart. The classical agent cannot distinguish the two physical realities because the input state is "informationally incomplete" in the above example. This is like the situation where one hopes to perform quantum state tomography with computational basis measurements alone — it is impossible because the measurements are not informationally complete.

One may wonder whether it is possible to learn the intrinsic descriptions of all actions if the classical agent do not have any prior knowledge about each action. Our first result shows that even without knowing what each action is, a classical agent can learn the intrinsic descriptions of all actions when the actions enable the agent to explore the quantum state space completely, and there is at least one action that implements a nontrivial POVM that depends on the state being measured.

Definition 1. A world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ is universal if there is $x \in \mathcal{X}$, such that the state ρ_x is pure; there are $y_1, \ldots, y_k \in \mathcal{Y}$, such that the actions maps $\mathcal{E}_{y_1}, \ldots \mathcal{E}_{y_k}$ constitute a universal set of unitary transformations; and there is $z \in \mathcal{Z}$, such that the POVM $\mathcal{M}_z$ has at least one POVM element not proportional to identity.

Given an arbitrary world model $\mathcal{W}$ that satisfies Definition 1, the classical agent would not know the action x that prepares a pure state, the actions $y_1, \ldots, y_k$ that implements the universal set of unitaries, nor the action z that implements the nontrivial POVM. Furthermore, the classical agent has no knowledge about what the physical operations $\rho_x, \mathcal{E}_{y_1}, \ldots \mathcal{E}_{y_k}, \mathcal{M}_z$ are. However, it is possible to learn all actions in $\mathcal{W}$.

Theorem 2 (Learning intrinsic description)**.** *Consider a (uncountably large) set $\mathcal{Q}^*$ of candidate world models that satisfy Definition 1. If the true world model $\mathcal{W}_{\text{true}} \in \mathcal{Q}^*$, then the classical agent can learn the description of every action in $\mathcal{W}_{\text{true}}$ to arbitrarily small error (up to one global unitary or anti-unitary transformation).*

Proof idea. Here, we present the general idea of the proof. The detail is given in Appendix D. First, we design a procedure for testing whether a composed CPTP map $\mathcal{E}_{y_k} \circ \ldots \circ \mathcal{E}_{y_1}$ is an identity. Then, we use the fact that unitaries are the only reversible CPTP maps and the identity testing procedure to create a protocol for testing whether a CPTP map $\mathcal{E}_y$ is a unitary. The unitary test allows the learning agent to identify every action that implements a unitary transformation.

Next, we show that randomly composing the identified unitaries forms an approximate Haar random unitary. This is proven using a contraction theorem for random walks on compact semi-simple Lie groups [47]. This result holds for any universal set of unitaries without the need to include inverses or have algebraic entries, but the theorem is weaker than the spectral gap theorem in [5]. Building on Corollary 7 in [47], we can show that the expectation value of any Lipschitz continuous function on a randomly composed unitaries is approximately equal to that of a Haar random unitary. Hence, although the learning agent doesn't know what each unitary is, the learning agent can still sample approximate Haar-random unitaries. In fact, it will suffice to sample approximately from a unitary two-design. Up to this point, the learning agent has not learned the description for any of the actions.

We then prove the following: given a procedure that samples approximately from a unitary two-design and the availability of an unknown POVM with at least one POVM element not proportional to identity, we can create a procedure that estimates the overlap $\text{tr}(\rho_1\rho_2)$ for any two states ρ_1, ρ_2. This procedure makes use of the two-design property of the random unitary ensemble.

Applying this overlap estimation procedure, the learning agent can determine for which value of x the state ρ_x is pure, and then reach other pure states by applying unitary circuits to ρ_x. Through further applications of the overlap estimation procedure, the learning agent can find a special set of states $\{|\psi_k\rangle\}$ with a particular geometry. Specifically, the special set of states corresponds to an orthonormal basis of pure states $|e_1\rangle, \ldots, |e_d\rangle$, superpositions of pairs of these basis states $\frac{1}{\sqrt{2}}(|e_i\rangle + |e_j\rangle)$, $\frac{1}{\sqrt{2}}(|e_i\rangle + \mathrm{i}\,|e_j\rangle)$ with a real or imaginary relative phase, and also superpositions of three basis states $\frac{1}{\sqrt{3}}(|e_1\rangle + |e_i\rangle + |e_j\rangle)$, $\frac{1}{\sqrt{3}}(|e_1\rangle + \mathrm{i}\,|e_2\rangle + \mathrm{i}\,|e_j\rangle)$, $\frac{1}{\sqrt{3}}(|e_1\rangle + |e_i\rangle + \mathrm{i}\,|e_j\rangle)$. The superposition of three bases is required to transfer information about the relative phases across different pairs of bases. The only ambiguity in this procedure is that all experiments would yield the same results if each state $|\psi_k\rangle$ were replaced by $U\,|\psi_k\rangle$ where U is a fixed unitary transformation, and/or if (i) were replaced by $(-i)$ in the superpositions of basis states (i.e., if a fixed antiunitary transformation were applied to each state $|\psi_k\rangle$).

Given the special set of states $\{|\psi_k\rangle\}$ and the procedure for estimating state overlaps, the classical agent can perform a version of quantum state tomography to learn the physical representation of ρ_x for all $x \in \mathcal{X}$. Similarly, the classical agent can learn the CPTP map $\mathcal{E}_y$ for all $y \in \mathcal{Y}$ by performing quantum state tomography on the states $\mathcal{E}_y(|\psi_k\rangle\langle\psi_k|)$ for all special states $|\psi_k\rangle$. The classical agent can also learn the POVM $\mathcal{M}_z$ for all $z \in \mathcal{Z}$ from the outcome probability distribution when $\mathcal{M}_z$ is performed on the special states $\{|\psi_k\rangle\}$. Hence, the classical agent has learned all the actions. □

When we cannot explore the quantum state space completely, it may be impossible to learn the intrinsic description to arbitrarily small error even with an infinite number of experiments. We present basic results that relates the learnability between different model classes in Appendix E. We build on these basic results to prove the unlearnability of various classes of quantum systems in Appendix F and G. To illustrate the challenge, consider an n-qubit system with an initial state ρ_0, a set of CPTP maps $\mathcal{E}_y, y \in \mathcal{Y}$ where $\mathcal{E}_y(I) = I$, and a POVM measurement $\mathcal{M}_0$. It is not hard to see that no algorithm can distinguish between the following two physical reality.

1. The initial state is slightly depolarized, $\rho_0 = 0.9|0^n\rangle\langle 0^n| + 0.1(I/2^n)$. But the computational basis measurement is perfect, $\mathcal{M}_0 = \{|b\rangle\langle b|\}_{b\in\{0,1\}^n}$.

2. The initial state is perfect, $\rho_0 = |0^n\rangle\langle 0^n|$. But the computational basis measurement is slightly depolarized, $\mathcal{M}_0 = \{0.9|b\rangle\langle b| + 0.1(I/2^n)\}_{b\in\{0,1\}^n}$.

This is an immediate consequence of the fact that unitary commutes with the action of the depolarization noise. We will present a nontrivial example in Section II B and discuss the practical implications.

C. Predicting extrinsic behaviors of the world model

We can also consider a less restrictive viewpoint. After the classical agent finishes conducting experiments to learn about the world model, we ask the classical agents to predict the outcome probability distribution for experiments composed of L CPTP maps; each such experiment is specified by $x, y_1, \ldots, y_L, z$. We give a formal definition in Appendix I. The classical agent does not have to find a physical description of the true world model $\mathcal{W}_{\text{true}}$ and is free to use any approach to make accurate prediction. The classical agent could always learn the extrinsic behavior by simply running all the $|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|$ experiments. But that procedure is very inefficient. In fact, in the worst case, no algorithm can do better than this highly inefficient procedure up to a logarithmic factor.

Theorem 3 (Worst case complexity for predicting extrinsic behaviors)**.** *To predict the probability of each experimental outcome to ϵ-error for any experiment with L maps, the classical agent has to perform at least $\Omega(|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|/\epsilon^2)$ experiments in the worst case, and the classical agent can always achieve the task by running $\tilde{\mathcal{O}}(|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|/\epsilon^2)$ experiments*[1]*.*

To derive the lower bound, we construct world models that behave like a special kind of maze. The classical agent has to navigate through the maze by performing a specific sequence of actions. Whenever the classical agent makes the wrong action, he/she fails. We then combine with a proof technique used in [9, 30] to establish the stated lower bound. The detailed proof is given in Appendix I.1. One way to avoid the worst case complexity is to assume that we have found a set of composed states and POVM elements, such that both sets span all possible states one can generate by performing actions in the world model. The following theorem shows that in such cases an exponential speed-up in L over the worst case can be achieved. The full algorithm and proof is given in Appendix J.

Theorem 4 (Predicting extrinsic behavior)**.** *Suppose we have found a set of unknown linearly independent states and a set of unknown POVM elements composed from $\rho_x, \mathcal{E}_y, \mathcal{M}_z$, such that both sets span all the states one can prepare. Then, we can predict the probability of each experimental outcome to ϵ-error for any experiment with L maps after running $\tilde{\mathcal{O}}((|\mathcal{X}| + L^2|\mathcal{Y}| + |\mathcal{Z}|)/\epsilon^2)$ experiments.*

As we have seen in previous sections, there are examples where we cannot learn the true physical reality, but existing gate set tomography protocols [4, 7, 20, 39] can still learn something in those examples. The resolution here is that existing gate set tomography protocols should be understood as methods for learning the extrinsic behaviors of the world model. Based on this resolution, the algorithm used in the proof of Theorem 4 can be seen as a rigorous learning algorithm with provable prediction guarantee for gate set tomography. A detailed discussion of gate set tomography viewed from our mathematical theory is given in Appendix K.

II. Case studies

After providing the foundation for describing and understanding what we could learn from quantum experiments, we present three case studies and discuss their practical implications.

[1] $\tilde{\mathcal{O}}(\cdot)$ neglects the logarithmic factors.

A. Learning under gate-independent noise on Clifford gates

A common assumption used in the randomized benchmarking literature [26, 33–35, 41] is that the noise process on Clifford gates are independent of the specific gate. This assumption considers all elements in the Clifford group and is crucial to ensure randomized benchmarking could extract an accurate value for the average gate fidelity [40]. More precisely, when the classical agent performs a Clifford gate C, the CPTP map implemented is

$$\mathcal{E}_C(\rho) = \mathcal{N}(C\rho C^\dagger), \tag{4}$$

where $\mathcal{N}$ is a CPTP map close to the identity that does not depend on the gate C. We leave open the possibility of only requiring a gate-independent noise on a few Clifford gates. In the benchmarking literature, it is also common to assume that there is a noisy zero state ρ_0 that is close to the all zero state $|0^n\rangle$, and there is a noisy computational basis measurement $\mathcal{M}_0 = \{M_b\}_{b\in\{0,1\}^n}$, where M_b is close to the state $|b\rangle\langle b|$. We will refer to these assumptions as bounded gate-independent noise on Clifford gates, and bounded noise on initial zero state and computational basis measurement. Typically, these assumptions are expected to hold only for a subsystem consisting of a constant number of qubits in a many-qubit quantum computer.

Under these assumptions, we give a simple and practical algorithm for learning every physical operation up to an unknown parameter $f = \langle 0^n|\,\rho_0\,|0^n\rangle$. The parameter f corresponds to the fidelity of the noisy zero state ρ_0, which is assumed to be close to one. Using techniques presented in the appendices, it is straightforward to show that f is unlearnable. To see this, assume Clifford gate noise $\mathcal{N}$ is identity, and use the result in Appendix G showing that one cannot distinguish between whether the state or the measurement is subject to a depolarizing channel. When Clifford gate noise $\mathcal{N}$ is not assumed to be identity, the learning does not become easier, hence f is still unlearnable.

When the Clifford gate noise $\mathcal{N}$ satisfies $\mathcal{N}(I) = I$, $\mathcal{N}$ will not depend on the unlearnable parameter f and can be learned up to arbitrarily small error. Under this further assumption, several existing algorithms robust to state preparation and measurement error have been proposed to learn $\mathcal{N}$ using information obtained from randomized benchmarking [26, 33, 41]. The best existing algorithm [26] learns $\mathcal{N}$ using $\mathcal{O}(d^8)$ experiments, where $d = 2^n$. Here, we show that the proposed algorithm only requires $\mathcal{O}(d^4 \log d)$ experiments. The number of experiments scale linearly in the number of parameters d^4 up to a logarithmic factor. The rigorous guarantees are presented in the following theorem and the proof is given in Appendix H.

Theorem 5 (Learning under gate-independent noise on Clifford gates)**.** *Assume bounded gate-independent noise on Clifford gates, and bounded noise on initial zero state and computational basis measurement. All states ρ_x, processes $\mathcal{E}_y$, and POVM measurements $\mathcal{M}_z$ in the quantum system can be learned up to a single unlearnable parameter $f = \langle 0^n|\,\rho_0\,|0^n\rangle$.*

Furthermore, the Choi matrix for the Clifford gate noise $\mathcal{N}$ can be learned up to ϵ error in Hilbert-Schmidt norm using $\mathcal{O}(d^4 \log d/\epsilon^2)$ experiments assuming $\mathcal{N}(I) = I$.

Proof idea. We conduct two sets of randomized experiments to learn about the noisy initial state ρ_0, the Clifford gate noise $\mathcal{N}$, and the noisy measurement $\mathcal{M}_0$. The first set of experiments prepares ρ_0, evolves by $\mathcal{E}_C$ for a random Clifford C, and measures $\mathcal{M}_0$. The second set of experiments prepares ρ_0, evolves by $\mathcal{E}_{C_1}$ for a random Clifford C_1, evolves by $\mathcal{E}_{C_2}$ for a second random Clifford C_2, and measures $\mathcal{M}_0$. The postprocessing is essentially the same as classical shadow formalism [29], which is based on least-square estimator for the quantum objects [22]. Despite the noise in state, maps, and measurements, the unitary 2-design property of random Clifford gates allows us to essentially neglect the noise processes and proceed with the same learning algorithm. We show that from the first set of experiments, we can learn the state ρ_0 and the POVM $\mathcal{N}^\dagger(\mathcal{M}_0)$, which corresponds applying $\mathcal{N}$ followed by the POVM $\mathcal{M}_0$, up to the unlearnable parameter f. Then, the second set of experiments allows us to learn $\mathcal{N}$, which can then be used to error correct the POVM $\mathcal{N}^\dagger(\mathcal{M}_0)$ to learn $\mathcal{M}_0$.

With data obtained from the first and the second sets of randomized experiments, we can learn all physical operations. To learn state ρ_x, we repeat the randomized experiment: prepare ρ_x, evolve by

$\mathcal{E}_C$ for a random Clifford C, and measures $\mathcal{M}_0$. To learn CPTP map $\mathcal{E}_y$, we repeat the randomized experiment: prepare ρ_0, evolve by $\mathcal{E}_{C_1}$ for a random Clifford C_1, evolve by $\mathcal{E}_y$, evolve by $\mathcal{E}_{C_2}$ for a second random Clifford C_2, and measures $\mathcal{M}_z$. To learn POVM $\mathcal{M}_z$, we repeat the randomized experiment: prepare ρ_0, evolve by $\mathcal{E}_C$ for a random Clifford C, and measures $\mathcal{M}_z$.

Unitary design property of random Clifford gates and standard concentration inequality can be used to characterize the number of required experiments. As an example, we show that the proposed algorithm learns the Clifford gate noise $\mathcal{N}$ to ϵ error from a total of $\mathcal{O}(d^4 \log d/\epsilon^2)$ randomized experiments assuming $\mathcal{N}(I) = I$. □

B. Bit-flip or phase-flip error?

In this case study, we consider the task of learning the noise on Clifford and T gates when the noise channel depends on the gate. Using randomized compiling for tailoring noise process [48], we can consider the CPTP map for a Clifford or T gate G to be given by

$$\mathcal{E}_G(\rho) = \mathcal{P}_G(G\rho G^\dagger), \tag{5}$$

where $\mathcal{P}_G$ is a Pauli channel that depends on the gate G and is close to the identity. However, in the following theorem, we show that it is impossible to learn the true gate-dependent Pauli noise even in a single-qubit system.

Theorem 6 (Gate-dependent Pauli noise is unlearnable with Clifford+T gates)**.** *Consider a qubit system. Suppose one can prepare $|0\rangle$ perfectly and any state ρ subject to a small noise, measure in the computational basis perfectly, and apply Clifford gates and T gate, where each gate is followed by an unknown gate-dependent Pauli noise channel close to identity. It is impossible for any algorithm to learn the gate-dependent Pauli noise channels to arbitrarily small error.*

Proof idea. The theorem is established by proving that one is unable to distinguish if a bit-flip (X) error is more likely to happen than a phase-flip (Z) error after the Hadamard gate, which correspond to two different physical realities not related by a global unitary or anti-unitary transformation. Note that the quantum system can only prepare zero state $|0\rangle$ and measure in the computational basis perfectly, which are tomographically incomplete to learn any quantum channel. Furthermore for a given gate G, we need to first learn the Pauli noise on the other gates in order to learn the Pauli noise $\mathcal{P}_G$ on G. This interdependency can not be resolved due to a graphical invariant in the actions of Clifford and T gates on Pauli operators. Together, one could use the reduction techniques in Appendix E to prove the unlearnability. The full proof is given in Appendix F. □

This result may seem to contradict claims in known protocols, such as gate set tomography [4, 7, 20, 39] and ACES [16]. To resolve the conflict, recall that gate set tomography learns an effective model that can only be used to predict the extrinsic behaviors. The effective model does not describe the intrinsic descriptions. Hence, gate set tomography is not learning the true noise process. In ACES [16], it was assumed that one could prepare any tensor product of single-qubit stabilizer states perfectly. By making this additional assumption, one could avoid this no-go theorem.

The main implication of this result applies to the setting where we would like to improve the quantum system by first identifying the noise process. For example, one may want to understand if the Hadamard gate is experiencing a higher rate of bit-flip error or phase-flip error, which could be a useful information for improving the device. Theorem 6 shows that sometimes it is impossible to identify the noise process by any benchmarking protocol.

C. Noisy quantum advantage in learning from experiments

Recently an experiment has been conducted on the Sycamore quantum processor that demonstrate a significant advantage in using quantum devices – a combination of quantum memory and quantum

computers – to learn about the physical systems and dynamics [28]. The experimental demonstration is based on a series of theoretical works [1, 9, 10, 30]. However, the existing theoretical works focused on an idealized setting, where the quantum device used for learning is perfectly known and controllable. The only unknown is the physical systems or dynamics we are interested in learning about. In practice, this is not the case. The quantum device is noisy, yet the experimental demonstrations on up to 40 qubits show that a significant quantum advantage still holds.

In this section, we establish provable advantage for the experiments conducted in [28], where the quantum device is noisy and the noise process is not known. We focus on the task of predicting many incompatible properties in physical systems studied in [9, 28, 30], which corresponds to the first experiment in [28]. In this task, the learning agent is given an unknown physical system described by a state ρ. In *conventional experiments*, the agent performs measurements on ρ to gather classical data. Based on the data, the agent can adaptively change the measurements to obtain more data. After multiple rounds of measurements, the agent learns a model of ρ. With the help of quantum devices, the agent can consider a more powerful form of experiments, which was refer to as quantum-enhanced experiments [28]. In *quantum-enhanced experiments*, the agent can load each copy of the state ρ into a quantum computer. After storing multiple copies of ρ in the quantum memory, the agent can perform quantum data analysis on multiple copies of ρ to learn a model of ρ. In both conventional and quantum-enhanced experiments, the agent use the learned model to predict properties of ρ.

Here, we compare conventional experiments with perfect measurements to quantum-enhanced experiments based on a noisy quantum device. The noisy quantum device can prepare a noisy and unknown initial product state, load a copy of the unknown physical system ρ into part of the qubits, apply noisy and unknown single- and two-qubit gates, as well as perform a noisy and unknown product measurement. We can apply multiple layers of gates to perform quantum processing. Each layer of gates contains multiple non-overlapping gates applied in parallel. The CPTP map implemented by each gate depends arbitrarily on all other gates implemented in the same layer. We *do not assume* that each gate is affected by only a constant number other gates. And the CPTP maps are *not* guaranteed to be close to the ideal unitary channel up to a constant error. Due to the lack of sufficient assumptions, it is impossible to learn the intrinsic descriptions of the device to arbitrary accuracy. Furthermore, an exponential number of experiments is required to learn the extrinsic behaviors to high accuracy because the noise processes are highly non-local. Because the noise cannot be learned efficiently, it is unclear how one can perform standard error mitigation techniques [14, 31, 46].

Even if the device cannot be learned to arbitrarily small error, we show that one can learn the intrinsic descriptions of the device up to a certain threshold. Recall that one of the central ideas in the proof of Theorem 2 for learning intrinsic descriptions is to find states that satisfy a particular geometry. The geometry fixes the intrinsic descriptions of the states, and hence we can learn the states. If we find states that approximately satisfy some of the geometry, then the intrinsic descriptions we learn will only be partially correct. Building on this idea, we give a procedure that finds operations such that the experimental outcomes satisfy a distinctive geometry. If the geometry is satisfied up to an error η, we can guarantee that the errors in each of the two-qubit operations is upper bounded by $\epsilon = \mathcal{O}(\eta)$. Details of the procedure and proof is given in Appendix L.2.

Despite having an unknown error of ϵ per two-qubit operations, Theorem 7 shows that a large polynomial advantage can be achieved with noisy quantum-enhanced experiments. The unknown state ρ does not need to have any quantum entanglement in the system to yield the quantum advantage.

Theorem 7 (Quantum advantage with noisy devices)**.** *There is a distribution over unentangled physical state ρ and properties we would like to predict, such that if we need N_{Q} samples of ρ in noisy quantum-enhanced experiments to predict the property, the required number of samples in noiseless conventional experiments must be $N_{\mathrm{C}} = \Omega(N_{\mathrm{Q}}^a)$, where $a = -\log(2)/(2\log(1-4\epsilon)) = \mathcal{O}(1/\epsilon)$.*

If each of the two-qubit operations has an error ϵ of at most 0.5%, then we can obtain a separation of $N_{\mathrm{C}} = \Omega(N_{\mathrm{Q}}^{17.15})$. The detailed proof of Theorem 7 is given in Appendix L.

III. Conclusion

This work provides a rigorous theory for reasoning about what we can learn from noisy quantum experiments. We established fundamental results to understand our ability to learn intrinsic physical descriptions and predict extrinsic behaviors. In a separate paper, we will present a series of practical applications for efficiently learning noisy quantum systems with many qubits based on the foundations established in this work.

The rigorous theory opens up several future directions. First of all, the essential mechanisms that determine the learnability and unlearnability of a class of world models are yet to be discovered. We have presented a set of deduction rules in Appendix E relating the learnability between different world model classes. Could we pinpoint a set of model classes where the learnability of all classes of world models could be derived using the set of deduction rules? Another interesting direction is to capture the central properties that enable humans to learn about the physical processes in cold atom systems, solid-state systems, and chemical systems, despite the unlearnability results. Could we incorporate these properties into a machine learning algorithm to enhance their learning ability?

On the practical side, how should we confront these unlearnability results? One way is to engineer the quantum system such that the noise in the system can be more easily learned. To proceed in this route, we need to have a better understanding on what model classes are easier to learn. Designing quantum systems with easier-to-learn operations can be related to a principle in software design, where one is advised to design programs that are easier to trouble-shoot. Answering these questions would not only help the development of quantum technology, but also shed light on what we could ultimately learn about the physical world.

Acknowledgments:

The authors thank Sitan Chen, Jordan Cotler, Jerry Li, Richard Kueng, and Thomas Vidick for valuable input and inspiring discussions. HH is supported by a Google PhD Fellowship. JP acknowledges funding from the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, (DE-NA0003525, DE-SC0020290), and the National Science Foundation (PHY-1733907). The Institute for Quantum Information and Matter is an NSF Physics Frontiers Center.

Appendices

Contents

A. Finite-dimensional quantum world models

We consider a general framework involving classical agents interacting with a finite-dimensional quantum world. This is a mathematical framework for reasoning about what quantum physicists could learn from a finite-dimensional quantum system that they can interact with by various means.

Definition A.1 (d-dimensional quantum world)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denoting the action space and a finite set $\mathcal{B}$ denoting the possible measurement outcomes. A d-dimensional quantum world $\mathcal{W}$ is a tuple with three sets

$$\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}}), \tag{A1}$$

where ρ_x is a d-dimensional density matrix, $\mathcal{E}_y$ is a d-dimensional CPTP map, $\mathcal{M}_z = \{M_{zb}\}_{b\in\mathcal{B}}$ is a POVM with finitely many elements indexed by $b \in \mathcal{B}$.

Consider $d = 2$, which is equivalent to a qubit system. We give an example to illustrate the above definition. Let $\mathcal{X} = \{(\theta, \phi) \mid \theta \in [0, \pi], \phi \in [0, 2\pi]\}$ be an uncountably large set. And we define

$$\rho_{\theta,\phi} = \frac{I + \sin(\theta)\cos(\phi)X + \sin(\theta)\sin(\phi)Y + \cos(\theta)Z}{2}. \tag{A2}$$

In this world, we can prepare any pure state on the single-qubit bloch sphere. Let $\mathcal{Y} = \{h, t\}$ be a finite set consisting of two elements. We consider $\mathcal{E}_h(\rho) = H\rho H^\dagger$ to be the Hadamard gate and

$\mathcal{E}_t(\rho) = T\rho T^\dagger$ to be the T gate. Finally, we consider $\mathcal{Z} = \{0\}$ to be a singleton and $\mathcal{B} = \{0, 1\}$ to be a two-outcome space, where $\mathcal{M}_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}$ is the computational basis measurement. $\mathcal{W}$ defines a single-qubit world where one can perform universal single-qubit quantum computation. Alternatively, one could also consider $\mathcal{X}$ to be a set of $\vec{n} \in \mathbb{R}^3$ with $\|\vec{n}\|_2 \leq 1$. Or we could consider $\mathcal{Y}$ to be a set of sequences where each sequence is a pulse sequence the experimentalist could use to control the qubit system. Intuitively, $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ contain descriptions of the actions an experimentalist could perform on the finite-dimensional quantum system, and $\mathcal{B}$ contains descriptions of the measurement outcomes.

Remark 1. In the above definition, we can have $x_1, x_2 \in \mathcal{X}$, such that $x_1 \neq x_2$ but $\rho_{x_1} = \rho_{x_2}$. This construction encodes the intuition that there could be two different actions an experimentalist could perform that will result in the same initial state. For example, we can generate the state $|1\rangle$ from $|0\rangle$ by applying π rotation along the X axis or the Y axis.

The experimentalists could interact with the d-dimensional quantum world by performing experiments. The experimentalist selects a state ρ_x to be prepared, composes various different evolutions $\mathcal{E}_{y_1}, \ldots, \mathcal{E}_{y_L}$, then reads out the final state through a chosen measurement apparatus $\mathcal{M}_z$.

Definition A.2 (Experiment)**.** Given a d-dimensional quantum world $\mathcal{W}$. An experiment is a list of finite elements given as

$$E = (x \in \mathcal{X}, y_1 \in \mathcal{Y}, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}). \tag{A3}$$

Each experiment results in an outcome $b \in \mathcal{B}$ with probability $\mathrm{tr}\left(M_{zb}\left(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1}\right)(\rho_x)\right)$.

As the experimentalists improve their physical control (lasers, cavity, etc.), more initial states ρ_x can be created, more evolutions $\mathcal{E}_y$ can be performed, and more types of qubit readout $\mathcal{M}_z$ can be achieved. We could imagine an ideal case, where we are able to generate all states, perform all operations, and conduct all measurements. We consider such a world model to be complete. A formal definition is given below.

Definition A.3 (Completeness)**.** A d-dimensional $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ is complete if

- for all states ρ, $\exists x \in \mathcal{X}$, $\rho_x = \rho$,
- for all CPTP maps $\mathcal{E}$, $\exists y \in \mathcal{Y}$, $\mathcal{E}_y = \mathcal{E}$,
- for all POVM $\mathcal{M}$ with outcomes indexed by $b \in \mathcal{B}$, $\exists z \in \mathcal{Z}$, $\mathcal{M}_z = \mathcal{M}$.

We say the world model $\mathcal{W}$ has been extended to a richer world model $\mathcal{W}'$ if $\mathcal{W}'$ contains more actions corresponding to more initial states, quantum evolutions, and POVM measurements. The formal definition is given below.

Definition A.4 (Extension)**.** A d-dimensional world model $\mathcal{W}' = \left(\{\rho'_x\}_{x\in\mathcal{X}'}, \{\mathcal{E}'_y\}_{y\in\mathcal{Y}'}, \{\mathcal{M}'_z\}_{z\in\mathcal{Z}'}\right)$ is an extension of a d-dimensional world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$, denoted as $\mathcal{W}' \rhd \mathcal{W}$, if the following conditions hold

- $\mathcal{X} \subseteq \mathcal{X}'$ and $\forall x \in \mathcal{X}, \rho_x = \rho'_x$ (State extension),
- $\mathcal{Y} \subseteq \mathcal{Y}'$ and $\forall y \in \mathcal{Y}, \mathcal{E}_y = \mathcal{E}'_y$ (CPTP map extension),
- $\mathcal{Z} \subseteq \mathcal{Z}'$ and $\forall z \in \mathcal{Z}, \mathcal{M}_z = \mathcal{M}'_z$ (POVM extension).

We are now ready to define equivalence between two world models. Before giving the formal definition, let us consider two 2-dimensional worlds $\mathcal{W}_A, \mathcal{W}_B$ with the same spaces $\mathcal{X} = \{0\}, \mathcal{Y} = \{h, t\}, \mathcal{Z} = \{0\}, \mathcal{B} = \{0, 1\}$. Furthermore, we consider the particular actions in the two world models $\mathcal{W}_A, \mathcal{W}_B$ to be given by

$$\rho_0^A = I/2, \qquad \mathcal{E}_h^A(\rho) = H\rho H^\dagger, \qquad \mathcal{E}_t^A(\rho) = T\rho T^\dagger, \qquad \mathcal{M}_0^A = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{A4}$$

$$\rho_0^B = I/2, \qquad \mathcal{E}_h^B(\rho) = I/2, \qquad \mathcal{E}_t^B(\rho) = I/2, \qquad \mathcal{M}_0^B = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}. \tag{A5}$$

It is not hard to show that we cannot distinguish between world A and B by performing experiments using the limited set of actions — in both cases the measurement outcomes are sampled from the uniform distribution. However, we can clearly see that the two world models are intrinsically different. In particular, in world A, the set of maps is a universal gate set that generates a dense subset of SU(2). But, in world B, all the maps are completely depolarizing channels. Even though world A and B can not be distinguish using the limited set of actions $\mathcal{X} = \{0\}, \mathcal{Y} = \{h, t\}, \mathcal{Z} = \{0\}$, the two worlds are fundamentally different. By adding new actions, such as the ability to prepare some non-trivial states, we can distinguish between the two world models by performing experiments.

To discuss these concepts in a rigorous manner, we formally define the following relations between two world models $\mathcal{W}_1, \mathcal{W}_2$. We consider two world models to be equal $\mathcal{W}_1 = \mathcal{W}_2$ if all of the states, maps, and POVMs are equal. And we say the two world models are different $\mathcal{W}_1 \neq \mathcal{W}_2$ if one of the states, maps, or POVMs is different. In the above example, the two world models $\mathcal{W}_A, \mathcal{W}_B$ are different because the CPTP maps are different $\mathcal{E}_h^A \neq \mathcal{E}_h^B$ and $\mathcal{E}_t^A \neq \mathcal{E}_t^B$.

Definition A.5 (Equality)**.** Consider two d-dimensional world models $\mathcal{W}_A, \mathcal{W}_B$ with the same spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. $\mathcal{W}_A$ is equal to $\mathcal{W}_B$, denoted as $\mathcal{W}_A = \mathcal{W}_B$, if all states are the same $\rho_x^A = \rho_x^B, \forall x \in \mathcal{X}$, all CPTP maps are the same $\mathcal{E}_y^A = \mathcal{E}_y^B, \forall y \in \mathcal{Y}$, and all POVMs are the same $M_{zb}^A = M_{zb}^B, \forall z \in \mathcal{Z}, b \in \mathcal{B}$.

Then, we consider two world models to be weakly indistinguishable if they can not be distinguished using the set of actions in the world model. In the example given in Equation (A4) and (A5), $\mathcal{W}_A, \mathcal{W}_B$ are weakly indistinguishable because the measurement outcomes are always uniformly distributed.

Definition A.6 (Weakly indistinguishable)**.** Consider two d-dimensional world models $\mathcal{W}_A, \mathcal{W}_B$ with the same spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. $\mathcal{W}_A$ and $\mathcal{W}_B$ are weakly indistinguishable if for any experiment $E = (x \in \mathcal{X}, y_1 \in \mathcal{Y}, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z})$ and outcome $b \in \mathcal{B}$, we have

$$\mathrm{tr}\left(M_{zb}^A \left(\mathcal{E}_{y_L}^A \circ \ldots \circ \mathcal{E}_{y_1}^A\right)(\rho_x^A)\right) = \mathrm{tr}\left(M_{zb}^B \left(\mathcal{E}_{y_L}^B \circ \ldots \circ \mathcal{E}_{y_1}^B\right)(\rho_x^B)\right), \tag{A6}$$

i.e., the probabilities for obtaining the outcome b in experiment E are the same.

And we say two world models are strongly indistinguishable or equivalent to one another if they can not be distinguished by adding any set of actions. World models $\mathcal{W}_A, \mathcal{W}_B$ are not equivalent because adding a non-completely-mixed state enables us to distinguish between $\mathcal{W}_A$ and $\mathcal{W}_B$.

Definition A.7 (Strongly indistinguishable / Equivalence)**.** Consider two d-dimensional world models $\mathcal{W}_A, \mathcal{W}_B$ with the same spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. $\mathcal{W}_A$ and $\mathcal{W}_B$ are equivalent or strongly indistinguishable, denoted as $\mathcal{W}_A \equiv \mathcal{W}_B$, if for all extensions of $\mathcal{W}_A$,

$$\mathcal{W}_A' = \left(\{\rho_x^{A'}\}_{x\in\mathcal{X}'}, \{\mathcal{E}_y^{A'}\}_{y\in\mathcal{Y}'}, \{\mathcal{M}_z^{A'}\}_{z\in\mathcal{Z}'}\right) \rhd \mathcal{W}_A, \tag{A7}$$

there exists an extension of $\mathcal{W}_B$ with the same action space $\mathcal{X}', \mathcal{Y}', \mathcal{Z}'$,

$$\mathcal{W}_B' = \left(\{\rho_x^{B'}\}_{x\in\mathcal{X}'}, \{\mathcal{E}_y^{B'}\}_{y\in\mathcal{Y}'}, \{\mathcal{M}_z^{B'}\}_{z\in\mathcal{Z}'}\right) \rhd \mathcal{W}_B, \tag{A8}$$

such that $\mathcal{W}_A'$ and $\mathcal{W}_B'$ are weakly indistinguishable.

The above definition of equivalence has a natural characterization given by Theorem A.8. Before stating the theorem, let us recall the definition of unitary and anti-unitary transformation U. Given a $d \times d$ complex matrix C with a chosen basis. We define $\overline{C}$ to be the matrix where we take complex conjugation for all entries in C. A unitary transformation U is a $d \times d$ unitary matrix with $U^{-1} = U^\dagger$ that transforms C to $UCU^{-1} = UCU^\dagger$. An anti-unitary transformation A is a product of a $d \times d$ unitary matrix U and the complex conjugation operator K that transforms C to $ACA^{-1} = U\overline{C}U^\dagger$. Theorem A.8 shows that equivalent world models are related by a unitary or anti-unitary transformation.

Theorem A.8 (A characterization of equivalence)**.** *Consider two d-dimensional world models with the same spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$,*

$$\mathcal{W}_A = \left(\{\rho_x^A\}_{x\in\mathcal{X}}, \{\mathcal{E}_y^A\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z^A\}_{z\in\mathcal{Z}}\right), \tag{A9}$$
$$\mathcal{W}_B = \left(\{\rho_x^B\}_{x\in\mathcal{X}}, \{\mathcal{E}_y^B\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z^B\}_{z\in\mathcal{Z}}\right). \tag{A10}$$

$\mathcal{W}_A \equiv \mathcal{W}_B$ if and only if there exists a unitary or anti-unitary transformation U, such that

$$\rho_x^B = U\rho_x^A U^{-1}, \qquad \forall x \in \mathcal{X}, \tag{A11}$$
$$\mathcal{E}_y^B(\cdot) = U\mathcal{E}_y^A(U^{-1}(\cdot)U)U^{-1}, \qquad \forall y \in \mathcal{Y}, \tag{A12}$$
$$M_{zb}^B = UM_{zb}^A U^{-1}, \qquad \forall z \in \mathcal{Z}, b \in \mathcal{B}. \tag{A13}$$

We defer the proof to Appendix B. As an example, the following two world models with the action spaces $\mathcal{X} = \{0\}, \mathcal{Y} = \{h, s\}$ for Hadamard and Phase gates, $\mathcal{Z} = \{0\}$, and outcome space $\mathcal{B} = \{0, 1\}$ are equivalent (related by an anti-unitary transformation XK):

$$\rho_0^A = |0\rangle\langle 0|, \quad \mathcal{E}_h^A(\rho) = H\rho H^\dagger, \quad \mathcal{E}_s^A(\rho) = S\rho S^\dagger, \quad \mathcal{M}_0^A = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{A14}$$
$$\rho_0^B = |1\rangle\langle 1|, \quad \mathcal{E}_h^B(\rho) = H\rho H^\dagger, \quad \mathcal{E}_s^B(\rho) = S\rho S^\dagger, \quad \mathcal{M}_0^B = \{|1\rangle\langle 1|, |0\rangle\langle 0|\}. \tag{A15}$$

The possibility to describe the same physical world by two distinct descriptions arises from the intrinsic degeneracy in quantum mechanics: the freedom to choose an arbitrary basis in the Hilbert space (the unitary relation) or reverse the direction of time (the anti-unitary relation).

B. A characterization of equivalence: Proof

We will focus on showing that $\mathcal{W}_A \equiv \mathcal{W}_B$ implies the existence of a unitary or anti-unitary transformation. The other direction can be shown easily by noting that for all world model extensions of $\mathcal{W}_A$, we can extend $\mathcal{W}_B$ using the same unitary or anti-unitary transformation U.

First, we extend world model $\mathcal{W}_A$ to world model $\mathcal{W}'_A$ that comes with an expanded state preparation actions $\mathcal{X}' = \mathcal{X} \cup \Omega^{\text{pure}}$ and an expanded measurement actions $\mathcal{Z}' = \mathcal{Z} \cup \Omega^{\text{pure-POVM}}$. In particular, $\rho_\xi^{A'}, \forall \xi \in \Omega^{\text{pure}}$ consists of all the d-dimensional pure states, and $\mathcal{M}_\zeta^{A'}, \forall \zeta \in \Omega^{\text{pure-POVM}}$ consists of all the POVMs such that a particular POVM element associated to $b^* \in \mathcal{B}$ is a pure state. The definition of equivalence shows that there exists an extension $\mathcal{W}'_B$ with the same action space as $\mathcal{W}'_A$ such that $\mathcal{W}'_A$ and $\mathcal{W}'_B$ are weakly indistinguishable, i.e., all experiments yield the same distribution. The above condition yields the following,

$$\text{tr}(M_{\zeta b^*}^{A'}\rho_\xi^{A'}) = \text{tr}(M_{\zeta b^*}^{B'}\rho_\xi^{B'}), \forall \xi \in \Omega^{\text{pure}}, \zeta \in \Omega^{\text{pure-POVM}}. \tag{B1}$$

This concludes the basic construction of the extended world models.

Recall that $M_{\zeta b^*}^{A'}, \forall \zeta \in \Omega^{\text{pure-POVM}}$ consists of all pure states. Hence, for each $\xi \in \Omega^{\text{pure}}$, there exists $\zeta(\xi) \in \Omega^{\text{pure-POVM}}$ such that $M_{\zeta b^*}^{A'} = \rho_\xi^{A'}$ are the same pure state. We extend $\rho_\xi^{A'}$ to an orthonormal set of basis consisting of d pure states $\{\rho_{\xi_1}^{A'}, \ldots, \rho_{\xi_d}^{A'}\}$, where $\xi_1 = \xi$. We have the following from the above construction and Eq. (B1),

$$\text{tr}\left(M_{\zeta(\xi_j)b^*}^{A'}\rho_{\xi_i}^{A'}\right) = \delta_{ij} = \text{tr}\left(M_{\zeta(\xi_j)b^*}^{B'}\rho_{\xi_i}^{B'}\right), \forall i, j \in \{1, \ldots, d\}, \tag{B2}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. We are now going to utilize the structure of the quantum states and POVM elements, in particular, a quantum state is a positive-semidefinite matrix with trace one, and a POVM element is a positive-semidefinite matrix with eigenvalues less than equal to one. In particular, we use the following basic lemma that can be established by induction.

Lemma B.1. *Consider $\ell \geq 1$. Given quantum states $\{\rho_i\}_{i=1,\ldots,\ell}$ and POVM elements $\{F_j\}_{j=1,\ldots,\ell}$. If $\text{tr}(F_j\rho_i) = \delta_{ij}, \forall i, j \in \{1, \ldots, \ell\}$, then the collection of eigenvectors of ρ_i with non-zero eigenvalues over all i from 1 to ℓ span a subspace with dimension at least ℓ.*

Proof. Consider the eigenvectors of ρ_i with non-zero eigenvalues to be $\{v_a^i\}_{a\in A_i}$ and the associated eigenvalues be $\{p_a^i\}_{a\in A_i}$, then $\sum_{a\in A_i} p_a^i (v_a^i)^\dagger F_i v_a^i = 1$. Since $0 \le (v_a^i)^\dagger F_i v_a^i \le 1$ (from the definition of POVM element), $p_a^i > 0$ (we only consider non-zero eigenvalues), $\sum_{a\in A_i} p_a^i = 1$ (from the definition of quantum state), and $\sum_{a\in A_i} p_a (v_a^i)^\dagger F_i v_a^i = 1$, we have

$$(v_a^i)^\dagger F_i v_a^i = 1, \forall a \in A_i. \tag{B3}$$

Similarly, for all $j \neq i$, $\sum_{a\in A_i} p_a (v_a^i)^\dagger F_j v_a^i = 0$ implies that $(v_a^i)^\dagger F_j v_a^i = 0, \forall a \in A_i$. Equivalently,

$$\sqrt{F_j} v_a^i = 0, \forall a \in A_i, \tag{B4}$$

when $i \neq j$. With the basic results given above, we are ready to prove the statement through induction. The base case $\ell = 1$ is trivially true. Suppose the statement holds for $\ell - 1$. We assume that all the eigenvectors of ρ_ℓ with non-zero eigenvalues lie in the span of the eigenvectors with non-zero eigenvalues for ρ_i with $i < \ell$. Under this assumption, for all $a \in A_\ell$, there exists a set of coefficients $\{c_{i,a'}\}$ such that $v_a^\ell = \sum_{i<\ell} \sum_{a'\in A_i} c_{i,a'} v_{a'}^i$. This implies that

$$1 = (v_a^\ell)^\dagger F_\ell v_a^\ell = \left(\sum_{i<\ell} \sum_{a'\in A_i} c_{i,a'} \sqrt{F_\ell} v_{a'}^i \right)^\dagger \left(\sum_{i<\ell} \sum_{a'\in A_i} c_{i,a'} \sqrt{F_\ell} v_{a'}^i \right) = 0. \tag{B5}$$

The first equality follows from Eq. (B3). The last equality follows from Eq. (B4). The contradiction shows that one of the eigenvectors of ρ_ℓ with non-zero eigenvalues is not in the span of the eigenvectors with non-zero eigenvalues for ρ_i with $i < \ell$. Therefore the statement holds for ℓ. □

If there exists $k \in \{1, \ldots, d\}$, such that the rank of $M^{B'}_{\zeta(\xi_k)b^*}$ is greater than one, then the $d-1$ states $\{\rho^{B'}_{\xi_i}\}_{i\neq k}$ must have their eigenvectors with non-zero eigenvalues span a $(d-2)$-dimensional subspace to ensure that $\mathrm{tr}\left(M^{B'}_{\zeta(\xi_k)b^*}\rho^{B'}_{\xi_i}\right) = 0$. However, from Lemma B.1, the eigenvectors of the $d-1$ states $\{\rho^{B'}_{\xi_i}\}_{i\neq k}$ with non-zero eigenvalues must span at least a $(d-1)$-dimensional state space. The contradiction implies that $M^{B'}_{\zeta(\xi_j)b^*}, \forall j \in \{1, \ldots, d\}$ must all be rank-one matrices. The condition $\mathrm{tr}\left(M^{B'}_{\zeta(\xi_1)b^*}\rho^{B'}_{\xi_1}\right) = 1$ then implies that $\rho^{B'}_{\xi_1}$ must be a pure state and $M^{B'}_{\zeta(\xi_1)b^*} = \rho^{B'}_{\xi_1}$. We have now shown the following statement:

$$\forall \xi \in \Omega^{\mathrm{pure}},\ \rho^{B'}_{\xi} \text{ is a pure state and the POVM element } M^{B'}_{\zeta(\xi)b^*} = \rho^{B'}_{\xi}. \tag{B6}$$

An implication of this result is that $\forall \xi_1, \xi_2 \in \Omega^{\mathrm{pure}}$, we have

$$\mathrm{tr}\left(\rho^{A'}_{\xi_1}\rho^{A'}_{\xi_2}\right) = \mathrm{tr}\left(M^{A'}_{\zeta(\xi_1)b^*}\rho^{A'}_{\xi_2}\right) = \mathrm{tr}\left(M^{B'}_{\zeta(\xi_1)b^*}\rho^{B'}_{\xi_2}\right) = \mathrm{tr}\left(\rho^{B'}_{\xi_1}\rho^{B'}_{\xi_2}\right), \tag{B7}$$

where the second equation follows from Eq. (B1).

We can now construct a transformation T over pure state space by considering $T(\rho^{A'}_{\xi}) = \rho^{B'}_{\xi}, \forall \xi \in \Omega^{\mathrm{pure}}$. T is a transformation that takes pure states to pure states that satisfies

$$\mathrm{tr}\left(\rho^{A'}_{\xi_1}\rho^{A'}_{\xi_2}\right) = \mathrm{tr}\left(T(\rho^{A'}_{\xi_1})T(\rho^{A'}_{\xi_2})\right), \forall \xi_1, \xi_2 \in \Omega^{\mathrm{pure}}, \tag{B8}$$

as a result of Eq. (B7). Such a transformation T is also known as a symmetry transformation. By Wigner's theorem, T must take the following form:

$$T(\rho) = U\rho U^{-1}, \tag{B9}$$

where U is a unitary or an anti-unitary transformation. For a proof of Wigner's theorem, see Appendix A of Chapter 2 in *The Quantum Theory of Fields, Vol. 1*, by Weinberg. The above representation of the symmetry transformation T yields

$$\rho^{B'}_{\xi} = U\rho^{A'}_{\xi}U^{-1},\ \forall \xi \in \Omega^{\mathrm{pure}}. \tag{B10}$$

Using Eq. (B6), we also have the following relation for a subset of POVM elements,

$$M^{B'}_{\zeta(\xi)b^*} = U M^{A'}_{\zeta(\xi)b^*} U^{-1},\ \forall \xi \in \Omega^{\text{pure}}. \tag{B11}$$

Intuitively, we will now use the pure states $\rho_\xi, \forall \xi \in \Omega^{\text{pure}}$ to perform quantum POVM tomography. Then use the rank-one POVM elements $M_{\zeta(\xi)b^*}, \forall \xi \in \Omega^{\text{pure}}$ to perform quantum state tomography. Finally, we use both ρ_ξ and $M_{\zeta(\xi)b^*}$ to perform quantum process tomography. Together, we have established the statement of Theorem A.8.

a. POVMs: $\forall z \in \mathcal{Z}, b \in \mathcal{B}$, we have $\text{tr}(M^{A'}_{zb}\rho^{A'}_\xi) = \text{tr}(M^{B'}_{zb}\rho^{B'}_\xi) = \text{tr}(M^{B'}_{zb} U \rho^{A'}_\xi U^{-1}), \forall \xi \in \Omega^{\text{pure}}$. The first equality follows from the weak indistinguishability between $\mathcal{W}'_A$ and $\mathcal{W}'_B$. The second equality follows from Eq. (B10). We consider a measure μ_ξ over ξ such that $\rho^{A'}_\xi$ forms the Haar measure over the pure state space. The Haar integration formulas $\int d\mu |\psi\rangle\langle\psi| = I/d$ and $\int d\mu |\psi\rangle\langle\psi|^{\otimes 2} = (I + \text{SWAP})/(d(d+1))$ give us

$$\frac{\text{tr}(M^{A'}_{zb})}{d} = \int d\mu_\xi \,\text{tr}(M^{A'}_{zb}\rho^{A'}_\xi) = \int d\mu_\xi \,\text{tr}(M^{B'}_{zb} U \rho^{A'}_\xi U^{-1}) = \frac{\text{tr}(M^{B'}_{zb})}{d}, \tag{B12}$$

$$\frac{\text{tr}(M^{A'}_{zb})I + M^{A'}_{zb}}{d(d+1)} = \int d\mu_\xi \,\text{tr}(M^{A'}_{zb}\rho^{A'}_\xi)\rho^{A'}_\xi = \int d\mu_\xi \,\text{tr}(M^{B'}_{zb} U \rho^{A'}_\xi U^{-1})\rho^{A'}_\xi \tag{B13}$$

$$= \frac{\text{tr}(M^{B'}_{zb})I + U^{-1} M^{B'}_{zb} U}{d(d+1)}. \tag{B14}$$

Therefore, $\forall z \in \mathcal{Z}, b \in \mathcal{B}, M^B_{zb} = M^{B'}_{zb} = U M^{A'}_{zb} U^{-1} = U M^A_{zb} U^{-1}$ as stated in Theorem A.8.

b. States: $\forall x \in \mathcal{X}$, we have $\text{tr}(M^{A'}_{\zeta(\xi)b^*}\rho^{A'}_x) = \text{tr}(M^{B'}_{\zeta(\xi)b^*}\rho^{B'}_x) = \text{tr}(U M^{A'}_{\zeta(\xi)b^*} U^{-1}\rho^{B'}_x), \forall \xi \in \Omega^{\text{pure}}$. The first equality follows from the weak indistinguishability between $\mathcal{W}'_A$ and $\mathcal{W}'_B$. The second equality follows from Eq. (B11). We consider a measure μ_ξ over ξ such that $M^{A'}_{\zeta(\xi)b^*}$ forms the Haar measure over the pure state space. The Haar integration formulas give us

$$\frac{I + \rho^{A'}_x}{d(d+1)} = \int d\mu_\xi \,\text{tr}(M^{A'}_{\zeta(\xi)b^*}\rho^{A'}_x) M^{A'}_{\zeta(\xi)b^*} \tag{B15}$$

$$= \int d\mu_\xi \,\text{tr}(U M^{A'}_{\zeta(\xi)b^*} U^{-1}\rho^{B'}_x) M^{A'}_{\zeta(\xi)b^*} = \frac{I + U^{-1}\rho^{B'}_x U}{d(d+1)}. \tag{B16}$$

Therefore, $\forall x \in \mathcal{X}, \rho^B_x = \rho^{B'}_x = U\rho^{A'}_x U^{-1} = U\rho^A_x U^{-1}$ as stated in Theorem A.8.

c. CPTP maps: $\forall y \in \mathcal{Y}$, we have $\forall \xi_1, \xi_2 \in \Omega^{\text{pure}}$,

$$\text{tr}(M^{A'}_{\zeta(\xi_2)b^*}\mathcal{E}^{A'}_y(\rho^{A'}_{\xi_1})) = \text{tr}(M^{B'}_{\zeta(\xi_2)b^*}\mathcal{E}^{B'}_y(\rho^{B'}_{\xi_1})) = \text{tr}(U M^{A'}_{\zeta(\xi)b^*} U^{-1}\mathcal{E}^{B'}_y(U\rho^{A'}_{\xi_1}U^{-1})). \tag{B17}$$

From the same analysis for states, we have

$$\mathcal{E}^{A'}_y(\rho^{A'}_{\xi_1}) = U^{-1}\mathcal{E}^{B'}_y(U\rho^{A'}_{\xi_1}U^{-1})U. \tag{B18}$$

Because $\rho^{A'}_{\xi_1}$ can be any pure state, we have

$$U\mathcal{E}^A_y(U^{-1}(\cdot)U)U^{-1} = U\mathcal{E}^{A'}_y(U^{-1}(\cdot)U)U^{-1} = \mathcal{E}^{B'}_y(\cdot) = \mathcal{E}^B_y(\cdot) \tag{B19}$$

as stated in Theorem A.8.

C. Foundations for learning intrinsic descriptions

The goal of learning is to conduct experiments to gain knowledge about the actual world model among a collection of potential world models. Learning theory provides a formal language to study such an information gathering process. For learning about a quantum-mechanical world, we would

like to know what quantum-mechanical operations each action $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$ corresponds to assuming the true world model belongs to a specified set of world models.

One should think of the learning process as follows. A classical agent is given the premise that the true world model belongs to some set of world models. Then, the classical agent conducts experiments to learn what the true model is. The set of possible models is called the concept class (often when each model is a classical Boolean function), hypothesis class (often when each model is a function from space X to space Y), or model class in machine learning. We will use the less loaded word, model class, to refer to a set of potential world models. In many branches of mathematics, including learning theory [37], a class is a set of mathematical objects. We will follow the same convention here.

Definition C.1 (Model class for d-dimensional quantum world)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ denoting the action spaces and set $\mathcal{B}$ denoting the outcome space. A d-dimensional model class $\mathcal{Q}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ is a set $\{\mathcal{W}\}$ of d-dimensional quantum world models with the same spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$.

Recall that some world models are equivalent to one another (describe the same physical reality) while not being equal (the mathematical description looks nominally different), i.e., $\mathcal{W}_1 \neq \mathcal{W}_2$ but $\mathcal{W}_1 \equiv \mathcal{W}_2$. This is intrinsic to the description of quantum mechanics as we have shown in Theorem A.8. When a model class contains two world models that are nominally different but physically the same, we say the model class is redundant. A formal definition is given below.

Definition C.2 (Redundant model class)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. $\mathcal{Q}$ is redundant if $\exists\, \mathcal{W}_1 \neq \mathcal{W}_2 \in \mathcal{Q}$, $\mathcal{W}_1 \equiv \mathcal{W}_2$.

Another basic concept about model classes is that two model classes could be equivalent to one another as a result of the equivalence of world models. For example, if we have three equivalent world models $\mathcal{W} \equiv \mathcal{W}_1 \equiv \mathcal{W}_2$, then the model class $\mathcal{Q} = \{\mathcal{W}\}$ contains the same set of equivalent world models as $\tilde{\mathcal{Q}} = \{\mathcal{W}_1, \mathcal{W}_2\}$. Hence, we say the two model classes are equivalent. We give the formal definition of equivalent model classes below.

Definition C.3 (Equivalent model classes)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Model classes $\mathcal{Q}, \tilde{\mathcal{Q}}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ are equivalent if and only if $\forall \mathcal{W} \in \mathcal{Q}, \exists \tilde{\mathcal{W}} \in \tilde{\mathcal{Q}}, \mathcal{W} \equiv \tilde{\mathcal{W}}$, and $\forall \tilde{\mathcal{W}} \in \tilde{\mathcal{Q}}, \exists \mathcal{W} \in \mathcal{Q}, \mathcal{W} \equiv \tilde{\mathcal{W}}$.

A redundant model class is not preferred as the same physical reality is described by two different representations. However, the following basic proposition shows that a redundant model class is equivalent to a non-redundant model class. For example, if $\mathcal{Q} = \{\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3\}$, where $\mathcal{W}_1 \equiv \mathcal{W}_2$ and $\mathcal{W}_1 \not\equiv \mathcal{W}_3$, then we have $\mathcal{Q}$ is equivalent to $\mathcal{Q}' = \{\mathcal{W}_1, \mathcal{W}_3\}$.

Proposition C.4. *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. There exists a non-redundant model class $\tilde{\mathcal{Q}}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, such that $\tilde{\mathcal{Q}}$ is equivalent to $\mathcal{Q}$.*

Proof. We partition all world models in $\mathcal{Q}$ into equivalence classes, where all models in the same equivalence class are equivalent to one another, and those that are in different equivalence classes are not equivalent. We choose one representative from each equivalence class of world models in $\mathcal{Q}$. We define $\tilde{\mathcal{Q}}$ to be the set of the representatives. $\tilde{\mathcal{Q}}$ is equivalent to $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ is non-redundant. □

After defining and illustrating some basic properties of model classes, we consider the learnability of a model class. We say a model class is learnable if for any world model $\mathcal{W}$ in the model class, the classical agent can identify the physical operations to an arbitrarily small error up to a global unitary or anti-unitary transformation using a finite number of experiments. The unitary or anti-unitary transformation U is necessary as Theorem A.8 states that two world models related by U are equivalent and describe the same physical reality. The transformation U corresponds to a change of basis in the quantum Hilbert space and potentially followed by a complex conjugation operation.

Definition C.5 (Learnability of a model class)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ for d-dimensional world models over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. The model class $\mathcal{Q}$ is learnable if $\forall\, \epsilon, \delta > 0, \forall \mathcal{W} \in \mathcal{Q}$, there exists a unitary or anti-unitary transformation U,

- $\forall x \in \mathcal{X}$, with probability $\geq 1-\delta$, a classical agent can conduct finitely many experiments as defined in Def. A.2 to output $\tilde{\rho}_x$ satisfying $\left\|\rho_x - U\tilde{\rho}_x U^{-1}\right\|_1 \leq \epsilon$,
- $\forall y \in \mathcal{Y}$, with probability $\geq 1-\delta$, a classical agent can conduct finitely many experiments as defined in Def. A.2 to output $\tilde{\mathcal{E}}_y$ satisfying $\left\|\mathcal{E}_y - U\tilde{\mathcal{E}}_y(U^{-1}(\cdot)U)U^{-1}\right\|_\diamond \leq \epsilon$.
- $\forall z \in \mathcal{Z}, b \in \mathcal{B}$, with probability $\geq 1-\delta$, a classical agent can conduct finitely many experiments as defined in Def. A.2 to output $\tilde{M}_{zb}$ satisfying $\left\|M_{zb} - U\tilde{M}_{zb}U^{-1}\right\|_1 \leq \epsilon$.

In many scenarios, it is too much to ask for the ability to learn everything about a world model, i.e., all initial states, CPTP maps, and measurements. For example, we might only want to predict a property of one of the possible initial state, such as its purity. Predicting properties can often be significantly more efficient than learning the full description [29]. Furthermore, even if a model class $\mathcal{Q}$ is unlearnable according to the above definition, we may still be able to predict some properties.

Definition C.6 (Predictability of a property)**.** Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, a model class $\mathcal{Q} = \{\mathcal{W}\}$ for d-dimensional world models over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, and a function f that maps a world model $\mathcal{W}$ to a property represented by a value in $\mathbb{R}$. The property f is predictable in the model class $\mathcal{Q}$ if $\forall \epsilon, \delta > 0, \forall \mathcal{W} \in \mathcal{Q}$, with probability $\geq 1-\delta$, a classical agent can conduct finitely many experiments as defined in Def. A.2 to output $\tilde{o} \in \mathbb{R}$ satisfying $|f(\mathcal{W}) - \tilde{o}| \leq \epsilon$.

D. A general theorem for learning intrinsic physical descriptions

The goal of this appendix is to prove the following theorem. Here, we consider a model class $\mathcal{Q}$ such that for any candidate world model $\mathcal{W}$ in $\mathcal{Q}$, there exists an action to prepare a pure state, a set of actions for implementing a universal set of unitaries, and an action for implementing a nontrivial POVM. A trivial POVM produces a measurement outcome independent of the input state. The actions that satisfy the above conditions could be different for different candidate world model $\mathcal{W}$. The classical agent has no knowledge of what these actions are and what the corresponding physical operations are. Furthermore the model class $\mathcal{Q}$ could contain uncountably many candidate world models. The theorem states that even without knowing what any action is, the classical agent can learn the intrinsic descriptions of all actions when the actions enable the exploration of the entire quantum state space.

Theorem D.1 (Restatement of Theorem 2)**.** *Given finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Consider a d-dimensional model class $\mathcal{Q}$ over action spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and outcome space $\mathcal{B}$. Suppose that for all $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_x\}_{x\in\mathcal{Z}}) \in \mathcal{Q}$,*

- *$\exists x \in \mathcal{X}$, ρ_x is pure.*
- *$\exists y_1, \ldots, y_k \in \mathcal{Y}$, $\mathcal{E}_{y_1}, \ldots \mathcal{E}_{y_k}$ constitute a universal set of unitary transformations.*
- *$\exists z \in \mathcal{Z}$, $\mathcal{M}_z$ has at least one element not proportional to identity.*

Then, $\mathcal{Q}$ is learnable.

To prove Theorem D.1, we give a learning algorithm such that for all world model $\mathcal{W}$ in $\mathcal{Q}$, the algorithm learns a world model $\tilde{\mathcal{W}}$ that satisfies $\tilde{\mathcal{W}} \equiv \mathcal{W}$ approximately, i.e., all physical descriptions of the actions in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are related by a global unitary or anti-unitary transformation (see Theorem A.8). The approximation error can be made arbitrarily close to zero as the algorithm conducts more experiments. From Definition C.5 on learnability of model class, we have $\mathcal{Q}$ is learnable.

In the following, we present an important lemma used in the proof. Then, we separate each step of the learning algorithm to learn the actions in the world model into subsections. The proof will consider the dimension d to be a constant.

D.1. A lemma on generating Haar-random unitaries

The proof of Theorem D.1 relies on the following lemma about the generation of approximate Haar measure using a universal set of unitaries. Here, we say $U_1, \ldots, U_k$ forms a universal set of unitaries if the set $\mathcal{U} = \{U_1, \ldots, U_k, U_1^{-1}, \ldots, U_k^{-1}\}$ generates a dense subgroup of the special unitary group. We follow the standard terminology, where the subgroup generated by the set $\mathcal{U}$ is the group consisting of elements that can be written as a product of elements in $\mathcal{U}$.

Definition D.2 (Real-valued Lipschitz function)**.** A real-valued Lipschitz function ϕ over the special unitary group satisfies $|\phi(U) - \phi(V)| \leq \|U - V\|_F$ for all unitaries U, V.

Lemma D.3 (Random unitaries approximately form Haar measure)**.** *Given k unitaries $U_1, \ldots, U_k$ that forms a universal set of unitaries. For any $\epsilon > 0$ and any real-valued Lipschitz function ϕ over the special unitary group, there exists $L > 0$, such that $\forall \ell \geq L$,*

$$\left| \frac{1}{k^\ell} \sum_{i_1=1}^{k} \ldots \sum_{i_\ell=1}^{k} \phi(U_{i_1} \ldots U_{i_\ell}) - \int d\mu_{\text{Haar}}(U) \phi(U) \right| < \epsilon, \tag{D1}$$

where μ_{Haar} is the Haar measure (uniform distribution) over the special unitary group.

We prove Lemma D.3 using a theorem given in [47], which is a corollary of a spectral gap theorem regarding semi-simple compact connected Lie group. Before stating the theorem, we give a few definitions. Consider a semi-simple compact connected Lie group G endowed with the bi-invariant Riemannian metric. The bi-invariant Riemannian metric gives a distance $d(g, h), \forall g, h \in G$. We define $\text{Lip}(G)$ to be the set of functions $\{\phi\} : G \to \mathbb{R}$ such that $\forall \phi \in \text{Lip}(G)$, $\sup_{g \neq h \in G} \frac{|\phi(g) - \phi(h)|}{d(g,h)} < \infty$. For $\phi \in \text{Lip}(G)$, we consider $\|\phi\|_{\text{Lip}} = \sup_{g \neq h \in G} \frac{|\phi(g) - \phi(h)|}{d(g,h)}$. We also define μ_{Haar} to be the Haar measure over G. The root second moment of a function $\phi : G \to \mathbb{R}$ is given by $\|\phi\|_2 = \sqrt{\int |\phi(x)|^2 d\mu_{\text{Haar}}(x)}$. For a probability measure μ, consider $\tilde{\mu}$ to be the probability measure such that

$$\int f(x) d\tilde{\mu}(x) = \int f(x^{-1}) d\mu(x) \tag{D2}$$

for all continuous function f. One can think of $\tilde{\mu}$ as the probability distribution over the inverse of the probability measure μ. For example, a uniform distribution over $U_1, \ldots, U_k$ yields a uniform distribution over $U_1^{-1}, \ldots, U_k^{-1}$. Consider the convolution between two probability measure μ and ν to be a probability measure,

$$(\mu * \nu)(g) = \sum_{h \in G} \mu(gh^{-1}) \nu(h). \tag{D3}$$

One can interpret $\mu * \nu$ as the probability distribution of $h\ell$ when we sample h according to μ and ℓ according to ν. For example, a convolution of the uniform distribution over $U_1, \ldots, U_k$ with itself would be a uniform distribution over $U_i U_j, \forall i, j = 1, \ldots, k$. Also, for a probability measure μ, we consider the support of μ, denoted as $\text{supp}(\mu)$, to be the intersection of every set A such that $\mu(A^c) = 0$.

Theorem D.4 (Corollary 7 in [47])**.** *Let G be a semi-simple compact connected Lie group endowed with the bi-invariant Riemannian metric, and μ be a probability measure on G. If $\text{supp}(\tilde{\mu} * \mu)$ generates a dense subgroup in G, then $\forall \psi_A \in \text{Lip}(G)$ with $\|\psi_A\|_2 = 1$ and $\int \psi_A(x) d\mu_{\text{Haar}}(x) = 0$, we have*

$$\|\psi_B\|_2 < 1 - \frac{c}{\log^A(\|\psi_A\|_{\text{Lip}} + 2)}, \tag{D4}$$

where $\psi_B(g) = \int \psi_A(h^{-1} g) d\mu(h)$, $A \leq 2$ depends on G, and $c > 0$ depends only on μ.

We are now ready to prove Lemma D.3. We apply Theorem D.4 by considering G to be the special unitary group with a representation in the vector space of matrices, and consider the bi-invariant Riemannian metric to be the Euclidean metric in the matrix space. The distance $d(U,V)$ on G is the the length of the shortest path from U to V on the special unitary group. Because we consider a constant dimensional special unitary group, we have

$$\|U-V\|_F \leq d(U,V) \leq C\|U-V\|_F, \tag{D5}$$

for a constant $C \geq 1$, where $\|X\|_F = \sqrt{\mathrm{tr}(X^\dagger X)}$. Hence, any real-valued function ϕ satisfying $|\phi(U)-\phi(V)| \leq \|U-V\|_F$ for all unitaries U,V is in $\mathrm{Lip}(G)$ with the Lipschitz norm $\|\phi\|_{\mathrm{Lip}} \leq 1$. Because ϕ is Lipschitz continuous, $\|\phi\|_2 < \infty$.

Proof for Lemma D.3. For the edge case where ϕ is a constant function, the lemma trivially holds. Let μ be the uniform distribution over $U_1^{-1},\ldots,U_k^{-1}$. The probability distribution $\tilde{\mu}$ over the inverse of μ is the uniform distribution over $U_1,\ldots,U_k$. $\tilde{\mu}*\mu$ is the uniform distribution over $U_iU_j^{-1}, \forall i,j=1,\ldots,k$, and $\mathrm{supp}(\tilde{\mu}*\mu)$ is $\{U_iU_j^{-1}, \forall i,j=1,\ldots,k\}$, which generates a dense subgroup of the special unitary group. Because ϕ is not a constant function, we can let ψ_1 be

$$\psi_1(g) = \frac{\phi(g) - \int \phi(U) d\mu_{\mathrm{Haar}}(U)}{\left\|\phi - \int \phi(U) d\mu_{\mathrm{Haar}}(U)\right\|_2}, \ \forall g \in G, \tag{D6}$$

which satisfies $\|\psi_1\|_2 = 1$, $\int \psi_1(x) d\mu_{\mathrm{Haar}}(x) = 0$, and

$$\|\psi_1\|_{\mathrm{Lip}} \leq \left\|\phi - \int \phi(U) d\mu_{\mathrm{Haar}(U)}\right\|_2 \|\phi\|_{\mathrm{Lip}} \leq \left\|\phi - \int \phi(U) d\mu_{\mathrm{Haar}}(U)\right\|_2. \tag{D7}$$

For any $\ell > 1$, we define

$$\psi_\ell(g) = \frac{\frac{1}{k^\ell}\sum_{i_1=1}^k \cdots \sum_{i_\ell=1}^k \phi(U_{i_1}\ldots U_{i_\ell}g) - \int \phi d\mu_{\mathrm{Haar}}}{\left\|\phi - \int \phi(U) d\mu_{\mathrm{Haar}}\right\|_2}, \ \forall g \in G, \tag{D8}$$

which satisfies $\int \psi_\ell(x) d\mu_{\mathrm{Haar}}(x) = 0$ and

$$|\psi_\ell(g_1) - \psi_\ell(g_2)| \leq \int \left|\psi_{\ell-1}(h^{-1}g_1) - \psi_{\ell-1}(h^{-1}g_2)\right| d\mu(h) \tag{D9}$$

$$\leq \|\psi_{\ell-1}\|_{\mathrm{Lip}} \int d(h^{-1}g_1, h^{-1}g_2) d\mu(h) \leq \|\psi_{\ell-1}\|_{\mathrm{Lip}} d(g_1,g_2). \tag{D10}$$

Hence, $\|\psi_\ell\|_{\mathrm{Lip}} \leq \|\psi_{\ell-1}\|_{\mathrm{Lip}} \leq \|\psi_1\|_{\mathrm{Lip}}$ for any $\ell > 1$.

We can apply Theorem D.4 with the probability measure μ defined above, $\psi_A = \psi_{\ell-1}/\|\psi_{\ell-1}\|_2$, and $\psi_B = \psi_\ell/\|\psi_{\ell-1}\|_2$ to obtain

$$\|\psi_\ell\|_2 \leq \left(1 - \frac{c}{\log^A(\|\psi_{\ell-1}\|_{\mathrm{Lip}} / \|\psi_{\ell-1}\|_2 + 2)}\right) \|\psi_{\ell-1}\|_2 \tag{D11}$$

$$\leq \left(1 - \frac{c}{\log^A(\|\psi_1\|_{\mathrm{Lip}} / \|\psi_{\ell-1}\|_2 + 2)}\right) \|\psi_{\ell-1}\|_2. \tag{D12}$$

The second inequality follows from $\|\psi_\ell\|_{\mathrm{Lip}} \leq \|\psi_1\|_{\mathrm{Lip}}$. By solving the above recursive relation, given any $\tilde{\epsilon} > 0$, there exists a large enough L such that

$$\|\psi_\ell\|_2 \leq \tilde{\epsilon}, \ \forall \ell \geq L. \tag{D13}$$

From the definition of ψ_ℓ and the fact that ϕ is Lipschitz-continuous over the special unitary group, $\forall \epsilon > 0$, there exists $L > 0$, such that $\forall \ell \geq L$, we have

$$\left|\frac{1}{k^\ell}\sum_{i_1=1}^k \cdots \sum_{i_\ell=1}^k \phi(U_{i_1}\ldots U_{i_\ell}g) - \int \phi d\mu_{\mathrm{Haar}}\right| \leq \epsilon, \ \forall g \in G. \tag{D14}$$

By setting $g = I$, we conclude the proof of Lemma D.3. □

D.2. The precision parameter

Consider the precision parameter η to be a fixed number at the start of the learning algorithm. Each subsection below corresponds to a set of subroutines in the learning algorithm that depends on η. After completion of all subroutines, the algorithm restarts with $\eta \leftarrow \eta/2$.

D.3. Testing identity and unitarity

We first present the subroutine for finding all actions $y_1, \ldots, y_{k'}$ that corresponds to unitary transformations, which is presented in the following lemma.

Lemma D.5 (Unitary identification)**.** *For a sufficiently small η, there is a subroutine that returns $y_1, \ldots, y_{k'}$ such that $\mathcal{E}_{y_1}, \ldots, \mathcal{E}_{y_{k'}}$ are all the unitary transformations in $\{\mathcal{E}_y\}_{y\in\mathcal{Y}}$.*

Proof. Consider an arbitrary norm $\|\cdot\|$ over the space of maps over quantum states. Assume that we have a subroutine for estimating how close a composition of CPTP maps $\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_\ell}$ is to an identity under the norm $\|\cdot\|$. The subroutine will be presented in Lemma D.6. Recall that in the finite set $\mathcal{Y}$, there are some CPTP maps $\mathcal{E}_{y_1}, \ldots, \mathcal{E}_{y_k}$ that correspond to unitary transformations. We define

$$\mathcal{Y}_{\text{unitary}} = \left\{\mathcal{E}_y^\dagger \mathcal{E}_y = I | \forall y \in \mathcal{Y}\right\}. \tag{D15}$$

We now present a subroutine that returns $\mathcal{Y}_{\text{unitary}}$. The proof of this lemma relies on a basic geometric fact about unitary: the only CPTP maps with some CPTP maps as their inverse are unitaries.

For each $y \in \mathcal{Y}$, we consider composing $\mathcal{E}_y$ with $\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_{\ell-1}}$ for $y'_1, \ldots, y'_{\ell-1} \in \mathcal{Y}$ and $\ell \leq 1/\eta$. The subroutine returns all actions $y \in \mathcal{Y}$ such that

$$\min_{y'_1,\ldots,y'_{\ell-1}} \left\| \mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_{\ell-1}} \circ \mathcal{E}_y - I \right\| \leq \eta. \tag{D16}$$

If $\mathcal{E}_y$ is a unitary transformation, we can find some $y'_1, \ldots, y'_{\ell-1}$ such that $\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_{\ell-1}} \circ \mathcal{E}_y$ is arbitrarily close to the identity under $\|\cdot\|$ as η goes to zero. If $\mathcal{E}_y$ is not a unitary transformation, then for all $y'_1, \ldots, y'_{\ell-1}$ and $\eta > 0$, there is a lower bound to how close $\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_{\ell-1}} \circ \mathcal{E}_y$ could be to the identity under $\|\cdot\|$. As a result, when η becomes small enough, the set of actions returned by the learning algorithm will be equal to $\mathcal{Y}_{\text{unitary}}$. □

Lemma D.6 (Identity testing)**.** *For a sufficiently small η and any $\epsilon > 0$, there exists a norm $\|\cdot\|$ over the space of maps over quantum states and a subroutine that takes in $y'_1, \ldots, y'_\ell$ and returns an estimate for*

$$\left\| \mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_\ell} - I \right\| \tag{D17}$$

up to an additive error ϵ with high probability.

Proof. The central property is that an identity map $\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_\ell}$ satisfies

$$\text{tr}(M_{zb} \left((\mathcal{E}_{y_{2k}} \circ \ldots \circ \mathcal{E}_{y_{k+1}}) \circ (\mathcal{E}_{y'_1} \circ \ldots \circ \mathcal{E}_{y'_\ell}) \circ (\mathcal{E}_{y_k} \circ \ldots \circ \mathcal{E}_{y_1}) \right) (\rho_x)) \tag{D18}$$

$$= \text{tr}(M_{zb} \left((\mathcal{E}_{y_{2k}} \circ \ldots \circ \mathcal{E}_{y_{k+1}}) \circ (\mathcal{E}_{y_k} \circ \ldots \circ \mathcal{E}_{y_1}) \right) (\rho_x)), \tag{D19}$$

for all $x \in \mathcal{X}, y_1, \ldots y_{2k} \in \mathcal{Y} \cup \{\text{NULL}\}, z \in \mathcal{Z}, b \in \mathcal{B}$. We denote the first term as $h(x, y_1, \ldots, y_{2k}, z, b)$ and the second term as $h_0(x, y_1, \ldots, y_{2k}, z, b)$. Here, the action $y = \text{NULL}$ corresponds to not implementing the action in the experiment defined in Def. A.2. The learning algorithm considers all possible composition with $2k = \lceil 1/\eta \rceil$, where η is the precision parameter.

The learning algorithm can obtain estimates for $h(x, y_1, \ldots, y_{2k}, z, b)$ and $h_0(x, y_1, \ldots, y_{2k}, z, b)$ by running the corresponding experiments with K repetitions. By Hoeffding's inequality, with $K =$

$\mathcal{O}(\log(1/\delta)/\epsilon^2)$, the algorithm can estimate $h(x, y_1, \ldots, y_{2k}, z, b)$ and $h_0(x, y_1, \ldots, y_{2k}, z, b)$ to ϵ-error with probability at least $1-\delta$. We consider K to be large enough such that the algorithm outputs an estimate for the quantity A defined as

$$A = \max_{x\in\mathcal{X}, y_1, \ldots y_{2k}\in\mathcal{Y}\cup\{\text{NULL}\}, z\in\mathcal{Z}, b\in\mathcal{B}} |h(x, y_1, \ldots, y_{2k}, z, b) - h_0(x, y_1, \ldots, y_{2k}, z, b)| \tag{D20}$$

$$= \max_{x\in\mathcal{X}, y_1, \ldots y_{2k}\in\mathcal{Y}\cup\{\text{NULL}\}, z\in\mathcal{Z}, b\in\mathcal{B}} \tag{D21}$$

$$\left|\text{tr}(M_{zb}\left((\mathcal{E}_{y_{2k}}\circ\ldots\circ\mathcal{E}_{y_{k+1}})\circ(\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}-I)\circ(\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})\right)(\rho_x))\right|. \tag{D22}$$

up to ϵ-error with high probability. We can interpret A as a norm $\|\cdot\|$ over the space of maps over quantum states when η is sufficiently small,

$$A = \left\|\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}-I\right\|. \tag{D23}$$

Positive definiteness follows from the fact that A is zero when $\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}$ is an identity; and A must be greater than zero for a sufficiently small η if $\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}$ is not equal to an identity from Lemma D.7. The two other conditions, absolute homogeneity and subadditivity, both follow from the definition of A in Eq. (D22). □

Lemma D.7 (Characterization of A)**.** *For a sufficiently small η, we have*

$$A = \max_{x\in\mathcal{X}, y_1, \ldots y_{2k}\in\mathcal{Y}\cup\{\text{NULL}\}, z\in\mathcal{Z}, b\in\mathcal{B}} \tag{D24}$$

$$\left|\text{tr}(M_{zb}\left((\mathcal{E}_{y_{2k}}\circ\ldots\circ\mathcal{E}_{y_{k+1}})\circ(\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}-I)\circ(\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})\right)(\rho_x))\right| > 0 \tag{D25}$$

if $\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}$ is not equal to an identity.

Proof. This claim follows from the assumption that there exists a universal set of unitaries and a pure state in the action space. Hence, we can generate a pure state $(\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})(\rho_x)$ such that

$$(\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell})((\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})(\rho_x)) \neq (\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})(\rho_x). \tag{D26}$$

From the assumption, we can also find a POVM $\mathcal{M}_z$ such that one of the POVM element M_{zb} is not proportional to the identity. There always exists a unitary U_1 that diagonalizes the Hermitian matrix

$$D \equiv (\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell})((\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})(\rho_x)) - (\mathcal{E}_{y_k}\circ\ldots\circ\mathcal{E}_{y_1})(\rho_x), \tag{D27}$$

such that the eigenvalues $\{\lambda_i^D\}$ are sorted from a greater value to a smaller value. Also, there exists a unitary U_2 that diagonalizes M_{zb} with eigenvalues $\{\lambda_i^{M_{zb}}\}$ sorted from large to small. We have

$$\text{tr}(U_2 M_{zb} U_2^\dagger U_1 D U_1^\dagger) = \sum_{i=1}^d \lambda_i^{M_{zb}}\lambda_i^D. \tag{D28}$$

Because $D \neq 0$ and $\text{tr}(D) = 0$, the largest eigenvalue $\lambda_1^D > 0$, the smallest eigenvalue $\lambda_d^D < 0$, and $\sum_{i=1}^d \lambda_i^D = 0$. We can thus find an index $k \geq 2$ such that $\lambda_i^D > 0, \forall i < k$ and $\lambda_i^D \leq 0, \forall i \geq k$. Since $M_{zb} \succeq 0$ is not proportional to identity, we have the largest eigenvalue $\lambda_1^{M_{zb}} > \lambda_d^{M_{zb}} \geq 0$.

$$\sum_{i=1}^d \lambda_i^{M_{zb}}\lambda_i^D = \sum_{i=1}^d (\lambda_i^{M_{zb}} - \lambda_k^{M_{zb}})\lambda_i^D > 0. \tag{D29}$$

When the precision parameter η is smaller enough, $(\mathcal{E}_{y_{2k}}\circ\ldots\circ\mathcal{E}_{y_{k+1}})$ can approximate any unitary with a properly chosen $y_{k+1}, \ldots, y_{2k}$ because there exists a universal set of unitaries by choosing the proper actions. Hence, there exists $x \in \mathcal{X}, y_1, \ldots y_{2k} \in \mathcal{Y}\cup\{\text{NULL}\}, z \in \mathcal{Z}, b \in \mathcal{B}$ such that $h(x, y_1, \ldots, y_{2k}, z, b) > h_0(x, y_1, \ldots, y_{2k}, z, b)$. Together, we see that A must be greater than zero if $\mathcal{E}_{y'_1}\circ\ldots\circ\mathcal{E}_{y'_\ell}$ is not equal to an identity. □

D.4. Estimating state overlaps

The learning algorithm has now identified a set $\{y_1, \ldots, y_{k'}\}$ of unitary transformation. The algorithm randomly composes the identified unitary transformations $\mathcal{E}_{y_1}, \ldots, \mathcal{E}_{y_{k'}}$. In particular, the algorithm randomly selects $1/\eta$ unitaries with replacement and compose them to form an approximate Haar random unitary. Using Lemma D.3, as η becomes smaller, we can obtain a better approximation to the Haar random unitary. The learning algorithm has no additional information other than the randomly composed operation is approximately Haar random.

The ability to generate Haar random unitary enables the learning algorithm to estimate state overlaps. Given two states ρ_1, ρ_2 which can be obtained by composing some initial states and CPTP maps, we have a randomized measurement procedure that guarantees the following.

Lemma D.8 (State overlap estimation with a fixed POVM)**.** *Given two states ρ_1, ρ_2, $\epsilon > 0$, $z \in \mathcal{Z}, b \in \mathcal{B}$, and a sufficiently small $\eta > 0$, there is a subroutine that estimates*

$$f_{M_{zb}}(\rho_1, \rho_2) = \alpha_{M_{zb}} \operatorname{tr}(\rho_1 \rho_2) + \beta_{M_{zb}} \tag{D30}$$

up to ϵ additive error, where $\alpha_{M_{zb}}, \beta_{M_{zb}}$ depends on POVM element M_{zb}.

Proof. Consider R repetitions. For repetition $r \in \{1, \ldots, R\}$, the subroutine performs:

1. Randomly compose $1/\eta$ actions in $\mathcal{Y}_{\text{unitary}}$ to generate a random CPTP map $\mathcal{E}$.
2. Measure the POVM $\mathcal{M}_z$ on $\mathcal{E}(\rho_1)$ and check if the measurement outcome is b.
3. Record a binary variable $C_r \in \{0, 1\}$ indicating if the outcome is b.
4. Measure the POVM $\mathcal{M}_z$ on $\mathcal{E}(\rho_2)$ and check if the measurement outcome is b.
5. Record a binary variable $D_r \in \{0, 1\}$ indicating if the outcome is b.

From Lemma D.9, we can show that $\hat{X} = \frac{1}{R} \sum_{r=1}^{R} C_r D_r$ is an accurate estimate for

$$\begin{aligned} f_{M_{zb}}(\rho_1, \rho_2) &= \frac{1}{d^2 - 1} \left((\operatorname{tr}(M_{zb})^2 - \operatorname{tr}(M_{zb}^2)/d) + \operatorname{tr}(\rho_1 \rho_2)(\operatorname{tr}(M_{zb}^2) - \operatorname{tr}(M_{zb})^2/d) \right), & \text{(D31)} \\ &= \alpha_{M_{zb}} \operatorname{tr}(\rho_1 \rho_2) + \beta_{M_{zb}}, & \text{(D32)} \end{aligned}$$

up to ϵ error when η is sufficiently small. Hence this lemma can be established. □

Lemma D.9 (Characterization of $\hat{X}$)**.** *Given two states ρ_1, ρ_2, $\epsilon > 0$, $z \in \mathcal{Z}, b \in \mathcal{B}$, a sufficiently large R and a sufficiently small $\eta > 0$, we have*

$$\left| \hat{X} - f_{M_{zb}}(\rho_1, \rho_2) \right| < \epsilon \tag{D33}$$

with high probability.

Proof. The expectation value of $\hat{X} = \frac{1}{R} \sum_{r=1}^{R} C_r D_r$ is equal to

$$\begin{aligned} \mathbb{E}\left[\frac{1}{R} \sum_{r=1}^{R} C_r D_r \right] &= \mathop{\mathbb{E}}_{\mathcal{E}} \operatorname{tr}(M_{zb} \mathcal{E}(\rho_1)) \operatorname{tr}(M_{zb} \mathcal{E}(\rho_2)) & \text{(D34)} \\ &\approx \int_U d\mu_{\text{Haar}}(U) \operatorname{tr}(M_{zb} U \rho_1 U^\dagger) \operatorname{tr}(M_{zb} U \rho_2 U^\dagger) & \text{(D35)} \\ &= \int_U d\mu_{\text{Haar}}(U) \operatorname{tr}((M_{zb} \otimes M_{zb})(U \otimes U)(\rho_1 \otimes \rho_2)(U^\dagger \otimes U^\dagger)) & \text{(D36)} \\ &= \frac{1}{d^2 - 1} \left((\operatorname{tr}(M_{zb})^2 - \operatorname{tr}(M_{zb}^2)/d) + \operatorname{tr}(\rho_1 \rho_2)(\operatorname{tr}(M_{zb}^2) - \operatorname{tr}(M_{zb})^2/d) \right). & \text{(D37)} \end{aligned}$$

Eq. (D35) is the consequence of the fact that a random composition of universal set of unitaries approximately forms a Haar random unitary. In particular, using the Lipschitz continuity of the function $\phi(U) \equiv \operatorname{tr}(M_{zb}U\rho_1U^\dagger)\operatorname{tr}(M_{zb}U\rho_2U^\dagger)$, Lemma D.3 shows that the approximation error can be made arbitrarily small as the number of composed unitary becomes sufficiently large (i.e., η sufficiently small). Because C_rD_r is a random variable bounded by one, using Hoeffding's inequality, we can choose $R = \mathcal{O}(\log(1/\delta))/\epsilon^2$, such that $\frac{1}{R}\sum_{r=1}^R C_rD_r$ equals to $\mathbb{E}_{\mathcal{E}}\operatorname{tr}(M_{zb}\mathcal{E}(\rho_1))\operatorname{tr}(M_{zb}\mathcal{E}(\rho_2))$ up to error $\epsilon/2$ with probability at least $1-\delta$. Eq. (D37), on the other hand, uses the second moment Haar integration formula over special unitary group. In particular, for $\mathrm{SU}(d)$ and $X \in \mathbb{C}^{(d\times d)\times(d\times d)}$, we have

$$\int_U d\mu_{\mathrm{Haar}}(U)(U\otimes U)X(U^\dagger\otimes U^\dagger) = \frac{1}{d^2-1}\left(I\operatorname{tr}(X) + S\operatorname{tr}(SX) - \frac{1}{d}S\operatorname{tr}(X) - \frac{1}{d}I\operatorname{tr}(SX)\right), \quad \text{(D38)}$$

where S is the swap operator over the tensor product space. Hence, when the precision parameter η is small enough and the number R of randomized experiments is large enough, $\hat{X} = \frac{1}{R}\sum_{r=1}^R C_rD_r$ is an accurate estimate for

$$f_{M_{zb}}(\rho_1,\rho_2) = \frac{1}{d^2-1}\left((\operatorname{tr}(M_{zb})^2 - \operatorname{tr}(M_{zb}^2)/d) + \operatorname{tr}(\rho_1\rho_2)(\operatorname{tr}(M_{zb}^2) - \operatorname{tr}(M_{zb})^2/d)\right), \quad \text{(D39)}$$

with an additive error at most ϵ. This establishes the claim. □

We are now ready to combine the two lemmas above to establish the main result of this subsection.

Lemma D.10 (State overlap estimation)**.** *Given two states ρ_1,ρ_2, $\epsilon > 0$, a sufficiently small $\eta > 0$, and the existence of a non-identity M_{zb} for some $z \in \mathcal{Z}, b \in \mathcal{B}$. There is a subroutine that estimates $\operatorname{tr}(\rho_1\rho_2)$ up to ϵ additive error.*

Proof. The learning algorithm utilizes the procedure in Lemma D.8 to build the subroutine achieving the claim of this lemma. Using the fact that for a d-dimensional vector x, $d\|x\|_2^2 \geq \|x\|_1^2$, we have

$$\operatorname{tr}(M_{zb}^2) - \operatorname{tr}(M_{zb})^2/d \geq 0. \quad \text{(D40)}$$

Furthermore, equality holds in Eq. (D40) if and only if all eigenvalues of M_{zb} are equal, which implies that M_{zb} is proportional to identity. If M_{zb} is proportional to identity, $f_{M_{zb}}(\rho_1,\rho_2)$ will be a constant function independent of ρ_1,ρ_2. In contrast, if M_{zb} is not proportional to identity, then for some $x \in \mathcal{X}, y \in \mathcal{Y}$, $f_{M_{zb}}(\rho_1,\rho_2)$ will be distinct between the following two pairs of states,

$$\rho_1 = \rho_x, \rho_2 = \mathcal{E}_y(\rho_x) \quad \text{and} \quad \rho_1 = \rho_x, \rho_2 = \rho_x. \quad \text{(D41)}$$

In particular, this is true if we choose the x such that ρ_x is pure and y such that $\mathcal{E}_y$ is one of the universal set of unitaries such that $\mathcal{E}_y(\rho_x) \neq \rho_x$.

From the assumption on the true world model (exists actions corresponding to preparation of a pure state, a universal set of unitaries, and a POVM element not proportional to identity), there always exists z,b,x,y such that $f_{M_{zb}}(\rho_1,\rho_2)$ is distinct under the two pairs of states in Eq. (D41). Hence, as η goes to zero, if M_{zb} is proportional to the identity, then the largest difference for the estimate of $f_{M_{zb}}(\rho_1,\rho_2)$ maximized over $x \in \mathcal{X}, y \in \mathcal{Y}$ will approach zero. In contrast, if M_{zb} is not proportional to the identity, the largest difference for the estimate of $f_{M_{zb}}(\rho_1,\rho_2)$ maximized over $x \in \mathcal{X}, y \in \mathcal{Y}$ will be greater than a positive value. Hence, we can consider an algorithm that finds the pair of $z \in \mathcal{Z}, b \in \mathcal{B}$ that yields the largest difference for the estimate of $f_{M_{zb}}(\rho_1,\rho_2)$ maximized over $x \in \mathcal{X}, y \in \mathcal{Y}$. The deduction above guarantees that the algorithm would find a POVM element M_{zb} that is not proportional to the identity under a sufficiently small η.

After finding a pair of z,b such that M_{zb} is not proportional to identity, we can now describe the procedure that estimates $\operatorname{tr}(\rho_1\rho_2)$ for any state ρ_1,ρ_2. Recall from Lemma D.8 and Eq. D40 that

$$f_{M_{zb}}(\rho_1,\rho_2) = \alpha_{M_{zb}}\operatorname{tr}(\rho_1\rho_2) + \beta_{M_{zb}}, \quad \text{(D42)}$$

where $\alpha_{M_{zb}} > 0$. Because $\alpha_{M_{zb}} > 0$, when $\rho_1 = \rho_2$, we can see that $f_{M_{zb}}(\rho_1,\rho_1)$ is maximized when ρ_1 is a pure state. The maximum value of $f_{M_{zb}}(\rho_1,\rho_2)$ is $\alpha_{M_{zb}} + \beta_{M_{zb}}$, and the minimum value is $\beta_{M_{zb}}$.

If ρ_1 is not a pure state, we can see that $f_{M_{zb}}(\rho_1, \rho_2) < \alpha_{M_{zb}} + \beta_{M_{zb}}$. The subroutine would hence go through all $x \in \mathcal{X}$ and find an x^* such that the estimate for $f_{M_{zb}}(\rho_x, \rho_x)$ is maximized. Recall from the assumption of the true world model, there exists an action x that prepares a pure state. The gap between the finite set of actions that prepare pure states and those that prepare mixed states allows us to guarantee that the action x^* we found prepares a pure state ρ_{x^*} when η is sufficiently small. Because ρ_{x^*} is a pure state, we have $\mathrm{tr}(\rho_{x^*}^2) = 1$ and hence from Eq. (D31),

$$f_{M_{zb}}(\rho_{x^*}, \rho_{x^*}) = \frac{1}{(d+1)d}(\mathrm{tr}(M_{zb}^2) + \mathrm{tr}(M_{zb})^2). \tag{D43}$$

The learning algorithm could obtain an estimate for The learning algorithm only has to determine $\mathrm{tr}(M_{zb})$ in order to determine $\mathrm{tr}(M_{zb}^2)$, which in turn fully specify the two values $\alpha_{M_{zb}}$ and $\beta_{M_{zb}}$.

An estimation for $\mathrm{tr}(M_{zb})$ can be obtained by reusing the randomized measurement data from the procedure described in Lemma D.8. We can show that $\hat{Y} = \frac{1}{R}\sum_{r=1}^{R} C_r$ is an accurate estimate for $\mathrm{tr}(M_{zb})/d$. Using the following first moment Haar integration formula over special unitary group,

$$\int_U d\mu_{\mathrm{Haar}}\, UXU^\dagger = \mathrm{tr}(X)I/d, \tag{D44}$$

and the standard concentration inequality, we have $\frac{1}{R}\sum_{r=1}^{R} C_r$ gives an estimate for $\frac{\mathrm{tr}(M_{zb})}{d}$ up to an error ϵ for sufficiently large R and sufficiently small η. Along with an estimate for Eq. (D43), the learning algorithm can determine both $\mathrm{tr}(M_{zb}^2)$ and $\mathrm{tr}(M_{zb})^2$, and hence $\alpha_{M_{zb}}$ and $\beta_{M_{zb}}$. Together, the learning algorithm can produce an accurate estimate for quantum state overlap $\mathrm{tr}(\rho_1\rho_2)$ from an estimate for $f_{M_{zb}}(\rho_1, \rho_2)$ given in Lemma D.8 and the estimates for $\alpha_{M_{zb}}$ and $\beta_{M_{zb}}$. This concludes the proof of this lemma. □

D.5. Learning descriptions of a special set of states

At this point, the algorithm still has not learned any description for any of the actions. However, the algorithm has identified several important actions. The algorithm has found $x^* \in \mathcal{X}$ where ρ_{x^*} is a pure state. The algorithm has also discovered $\mathcal{Y}_{\mathrm{unitary}} = \{y_1, \ldots, y_k\} \in \mathcal{Y}$ that forms a universal set of unitaries, which we will now denote as $U_{y_1}, \ldots, U_{y_k} \in \mathrm{SU}(d)$. Furthermore, the algorithm has now obtained a subroutine that provides accurate estimate for state overlap $\mathrm{tr}(\rho_1\rho_2)$. The learning algorithm can now utilize these tools to construct the entire structure of quantum state space. More precisely, the algorithm will find a special set of pure quantum states $\{\rho_i\}$ that satisfies a certain geometry. The algorithm generates the special set of pure states by applying compositions of the unitaries $U_y, \forall y \in \mathcal{Y}_{\mathrm{unitary}}$ onto the pure state ρ_{x^*}. We will limit the algorithm to consider a composition of length at most $1/\eta$. This means that the algorithm will only find a collection of states that satisfies the geometry *approximately*. However, an approximate geometry with small error implies that the learned descriptions will only be subject to a small error. As η goes to zero, the geometry and the learned description will become accurate to an arbitrarily small error. The geometry enables us to provide the intrinsic physical descriptions for states in the special set. Using properties of the geometry, we can guarantee that the description for the special set of pure states is accurate up to the equivalence relation – a global unitary or anti-unitary transformation – characterized by Theorem A.8. The construction of the special set of pure states is related to the proof of Wigner's theorem.

We denote the special collection of pure states as $\rho_i^{(\mathrm{basis})}, \forall i \in \{1, \ldots, d\}$, $\rho_{ij}^{(\mathrm{real})}, \forall i \neq j \in \{1, \ldots, d\}$, $\rho_{ij}^{(\mathrm{imag})}, \forall i \neq j \in \{1, \ldots, d\}$, $\rho_{ij}^{(\mathrm{triplet})}, \forall i \neq j \in \{2, \ldots, d\}$, $\rho_j^{(\mathrm{triplet,12})}, \forall j \in \{3, \ldots, d\}$ and $\rho_{ij}^{(\mathrm{triplet,i})}, \forall i \neq j \in \{2, \ldots, d\}$. The geometry of the states is given by the following equations.

$$\mathrm{tr}(\rho_i^{(\mathrm{basis})}\rho_j^{(\mathrm{basis})}) = \delta_{ij}, \quad \forall i, j \in \{1, \ldots, d\}, \quad \textit{(Fix the basis)} \tag{D45}$$

$$\mathrm{tr}(\rho_i^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{real})}) = \frac{1}{2}, \quad \forall i \neq j \in \{1, \ldots, d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{real})}\textit{)} \tag{D46}$$

$$\mathrm{tr}(\rho_j^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{real})}) = \frac{1}{2}, \quad \forall i \neq j \in \{1, \ldots, d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{real})}\textit{)} \tag{D47}$$

$$\mathrm{tr}(\rho_1^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet})}\textit{)} \tag{D48}$$

$$\mathrm{tr}(\rho_i^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet})}\textit{)} \tag{D49}$$

$$\mathrm{tr}(\rho_j^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet})}\textit{)} \tag{D50}$$

$$\mathrm{tr}(\rho_{1i}^{(\mathrm{real})}\rho_{ij}^{(\mathrm{triplet})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +1, a)} \tag{D51}$$

$$\mathrm{tr}(\rho_{1j}^{\mathrm{real}}\rho_{ij}^{(\mathrm{triplet})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +1, b)} \tag{D52}$$

$$\mathrm{tr}(\rho_{ij}^{(\mathrm{real})}\rho_{ij}^{(\mathrm{triplet})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +1, c)} \tag{D53}$$

$$\mathrm{tr}(\rho_i^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{imag})}) = \frac{1}{2}, \quad \forall i \neq j \in \{1,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{imag})}\textit{)} \tag{D54}$$

$$\mathrm{tr}(\rho_j^{\mathrm{basis}}\rho_{ij}^{(\mathrm{imag})}) = \frac{1}{2}, \quad \forall i \neq j \in \{1,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{imag})}\textit{)} \tag{D55}$$

$$\mathrm{tr}(\rho_{ij}^{\mathrm{real}}\rho_{ij}^{(\mathrm{imag})}) = \frac{1}{2}, \quad \forall i \neq j \in \{1,\ldots,d\}, \quad \textit{(Partially fix the phase for } \rho_{ij}^{(\mathrm{imag})}\textit{)} \tag{D56}$$

$$\mathrm{tr}(\rho_1^{(\mathrm{basis})}\rho_j^{(\mathrm{triplet},12)}) = \frac{1}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_j^{(\mathrm{triplet},12)}\textit{)} \tag{D57}$$

$$\mathrm{tr}(\rho_2^{(\mathrm{basis})}\rho_j^{(\mathrm{triplet},12)}) = \frac{1}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_j^{(\mathrm{triplet},12)}\textit{)} \tag{D58}$$

$$\mathrm{tr}(\rho_j^{(\mathrm{basis})}\rho_j^{(\mathrm{triplet},12)}) = \frac{1}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_j^{(\mathrm{triplet},12)}\textit{)} \tag{D59}$$

$$\mathrm{tr}(\rho_{2j}^{(\mathrm{real})}\rho_j^{(\mathrm{triplet},12)}) = \frac{2}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Transfer relative phase +1, a')} \tag{D60}$$

$$\mathrm{tr}(\rho_{12}^{(\mathrm{imag})}\rho_j^{(\mathrm{triplet},12)}) = \frac{2}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Transfer relative phase +i, b')} \tag{D61}$$

$$\mathrm{tr}(\rho_{1j}^{(\mathrm{imag})}\rho_j^{(\mathrm{triplet},12)}) = \frac{2}{3}, \quad \forall j \in \{3,\ldots,d\}, \quad \textit{(Transfer relative phase +i, c')} \tag{D62}$$

$$\mathrm{tr}(\rho_1^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet,i})}\textit{)} \tag{D63}$$

$$\mathrm{tr}(\rho_i^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet,i})}\textit{)} \tag{D64}$$

$$\mathrm{tr}(\rho_j^{(\mathrm{basis})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{1}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Fix absolute amplitude for } \rho_{ij}^{(\mathrm{triplet,i})}\textit{)} \tag{D65}$$

$$\mathrm{tr}(\rho_{1i}^{(\mathrm{real})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +1, a'')} \tag{D66}$$

$$\mathrm{tr}(\rho_{1j}^{(\mathrm{imag})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +i, b'')} \tag{D67}$$

$$\mathrm{tr}(\rho_{ij}^{(\mathrm{imag})}\rho_{ij}^{(\mathrm{triplet,i})}) = \frac{2}{3}, \quad \forall i \neq j \in \{2,\ldots,d\}, \quad \textit{(Transfer relative phase +i, c'').} \tag{D68}$$

We comment on each geometric constraint, so it would be easier to refer to in the following analysis. The geometry determines the description of the set of states.

Lemma D.11 (Geometry and states)**.** *The geometry of the states given in Eq.* (D45) *to Eq.* (D68) *is satisfied if and only if*

$$\rho_i^{(\mathrm{basis})} = U|i\rangle\langle i|U^{-1}, \qquad \forall i \in \{1,\ldots,d\}, \tag{D69}$$

$$\rho_{ij}^{(\mathrm{real})} = \frac{1}{2}U\left(|i\rangle + |j\rangle\right)\left(\langle i| + \langle j|\right)U^{-1}, \qquad \forall i \neq j \in \{1,\ldots,d\}, \tag{D70}$$

$$\rho_{ij}^{(\mathrm{imag})} = \frac{1}{2}U\left(|i\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle i| - \mathrm{i}\,\langle j|\right)U^{-1}, \qquad \forall i \neq j \in \{2,\ldots,d\}, \tag{D71}$$

$$\rho_{ij}^{(\mathrm{triplet})} = \frac{1}{3}U\left(|1\rangle + |i\rangle + |j\rangle\right)\left(\langle 1| + \langle i| + \langle j|\right)U^{-1}, \qquad \forall i \neq j \in \{2,\ldots,d\}, \tag{D72}$$

$$\rho_j^{(\text{triplet},12)} = \frac{1}{3}U\left(|1\rangle + \mathrm{i}\,|2\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle 1| - \mathrm{i}\,\langle 2| - \mathrm{i}\,\langle j|\right)U^{-1}, \qquad \forall j \in \{3,\ldots,d\}, \tag{D73}$$

$$\rho_{ij}^{(\text{triplet,i})} = \frac{1}{3}U\left(|1\rangle + |i\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle 1| + \langle i| - \mathrm{i}\,\langle j|\right)U^{-1}, \qquad \forall i \neq j \in \{2,\ldots,d\} \tag{D74}$$

for a unitary or anti-unitary transformation U.

Proof. One can directly verify that the set of states given in Eq. (D69) to Eq. (D74) satisfies the geometry given in Eq. (D45) to Eq. (D68). For the other direction, we utilize the following steps. The basic idea is to use the geometric constraints to gradually determine the descriptions for the states.

1. The constraint *(Fix the basis)* ensures that there exists a unitary transformation $U^{(0)}$ such that

$$(U^{(0)})^{-1}\rho_i^{(\text{basis})}(U^{(0)}) = |i\rangle\langle i|, \ \forall i \in \{1,\ldots,d\}. \tag{D75}$$

2. The constraint *(Fix absolute amplitude for $\rho_{ij}^{(\text{real})}$)* ensures that

$$(U^{(0)})^{-1}\rho_{ij}^{(\text{real})}(U^{(0)}) = \frac{1}{2}\left(|i\rangle + \mathrm{e}^{\mathrm{i}\phi_{ij}}\,|j\rangle\right)\left(\langle i| + \mathrm{e}^{-\mathrm{i}\phi_{ij}}\,\langle j|\right), \forall i \neq j \in \{1,\ldots,d\} \tag{D76}$$

for some unknown phase $\phi_{ij} \in [0, 2\pi)$. Consider $U^{(1)} = U^{(0)}D$, where D is a diagonal matrix with $D_{11} = 1$ and $D_{ii} = \mathrm{e}^{\mathrm{i}\phi_{1i}}, \forall i \in \{2,\ldots,d\}$. We have

$$(U^{(1)})^{-1}\rho_i^{(\text{basis})}(U^{(1)}) = |i\rangle\langle i|, \qquad \forall i \in \{1,\ldots,d\}, \tag{D77}$$

$$(U^{(1)})^{-1}\rho_{1i}^{(\text{real})}(U^{(1)}) = \frac{1}{2}\left(|1\rangle + |i\rangle\right)\left(\langle 1| + \langle i|\right), \qquad \forall i \in \{2,\ldots,d\}, \tag{D78}$$

$$(U^{(1)})^{-1}\rho_{ij}^{(\text{real})}(U^{(1)}) = \frac{1}{2}\left(|i\rangle + \mathrm{e}^{\mathrm{i}\phi'_{ij}}\,|j\rangle\right)\left(\langle i| + \mathrm{e}^{-\mathrm{i}\phi'_{ij}}\,\langle j|\right), \quad \forall i \geq 2, i \neq j \in \{1,\ldots,d\}, \tag{D79}$$

where $\phi'_{ij} \in [0, 2\pi)$ is some unknown phase.

3. From the constraints *(Fix absolute amplitude for $\rho_{ij}^{(\text{triplet})}$)*, *(Transfer relative phase $+1$, a)* and *(Transfer relative phase $+1$, b)* and Eq. (D78), we have

$$(U^{(1)})^{-1}\rho_{ij}^{(\text{triplet})}(U^{(1)}) = \frac{1}{3}\left(|1\rangle + |i\rangle + |j\rangle\right)\left(\langle 1| + \langle i| + \langle j|\right), \quad \forall i \neq j \in \{2,\ldots,d\}. \tag{D80}$$

4. The constraint *(Transfer relative phase $+1$, c)*, Eq. (D79) and Eq. (D80) ensure that

$$(U^{(1)})^{-1}\rho_{ij}^{(\text{real})}(U^{(1)}) = \frac{1}{2}\left(|i\rangle + |j\rangle\right)\left(\langle i| + \langle j|\right), \forall i \neq j \in \{1,\ldots,d\}. \tag{D81}$$

The state $\rho_{ij}^{(\text{triplet})}$ serves as an intermediate point to transfer relative phases.

5. The constraints *(Fix absolute amplitude for $\rho_{ij}^{(\text{imag})}$)*, *(Partially fix the phase for $\rho_{ij}^{(\text{imag})}$)*, and Eq. (D81) ensure that

$$(U^{(1)})^{-1}\rho_{ij}^{(\text{imag})}(U^{(1)}) = \frac{1}{2}\left(|i\rangle + s_{ij}\mathrm{i}\,|j\rangle\right)\left(\langle i| - s_{ij}\mathrm{i}\,\langle j|\right), \forall i \neq j \in \{1,\ldots,d\}, \tag{D82}$$

where $s_{ij} = \pm 1$ is an unknown phase. If $s_{12} = 1$, we define $U^{(2)} = U^{(1)}$. If $s_{12} = -1$, we define $U^{(2)} = U^{(1)}K$, where K is the complex conjugation operation. We have $U^{(2)}$ is either a unitary or anti-unitary transformation. Using the newly defined $U^{(2)}$, we have

$$(U^{(2)})^{-1}\rho_i^{(\text{basis})}(U^{(2)}) = |i\rangle\langle i|, \qquad \forall i \in \{1,\ldots,d\}, \tag{D83}$$

$$(U^{(2)})^{-1}\rho_{ij}^{(\text{real})}(U^{(2)}) = \frac{1}{2}\left(|i\rangle + |j\rangle\right)\left(\langle i| + \langle j|\right), \qquad \forall i \neq j \in \{1,\ldots,d\}, \tag{D84}$$

$$(U^{(1)})^{-1}\rho_{ij}^{(\mathrm{triplet})}(U^{(1)}) = \frac{1}{3}\left(|1\rangle + |i\rangle + |j\rangle\right)\left(\langle 1| + \langle i| + \langle j|\right), \qquad \forall i \neq j \in \{2,\ldots,d\}, \tag{D85}$$

$$(U^{(2)})^{-1}\rho_{12}^{(\mathrm{imag})}(U^{(2)}) = \frac{1}{2}\left(|1\rangle + \mathrm{i}\,|2\rangle\right)\left(\langle 1| - \mathrm{i}\,\langle 2|\right), \tag{D86}$$

$$(U^{(2)})^{-1}\rho_{ij}^{(\mathrm{imag})}(U^{(2)}) = \frac{1}{2}\left(|i\rangle + s'_{ij}\mathrm{i}\,|j\rangle\right)\left(\langle i| - s'_{ij}\mathrm{i}\,\langle j|\right), \qquad \forall i \neq j \in \{1,\ldots,d\}, \tag{D87}$$

for some $s'_{ij} \in \{\pm 1\}$.

6. From *(Fix absolute amplitude for $\rho_j^{(\mathrm{triplet},12)}$)*, *(Transfer relative phase $+1$, a')*, and *(Transfer relative phase $+\mathrm{i}$, b')*, we have

$$(U^{(2)})^{-1}\rho_{j}^{(\mathrm{triplet},12)}(U^{(2)}) = \frac{1}{3}\left(|1\rangle + \mathrm{i}\,|2\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle 1| - \mathrm{i}\,\langle 2| - \mathrm{i}\,\langle j|\right), \quad \forall j \in \{3,\ldots,d\}. \tag{D88}$$

7. From *(Transfer relative phase $+\mathrm{i}$, c')*, we have

$$(U^{(2)})^{-1}\rho_{1j}^{(\mathrm{imag})}(U^{(2)}) = \frac{1}{2}\left(|1\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle 1| - \mathrm{i}\,\langle j|\right), \quad \forall j \in \{2,\ldots,d\}, \tag{D89}$$

Similar to $\rho_{ij}^{(\mathrm{triplet})}$, the state $\rho_j^{(\mathrm{triplet},12)}$ serves as an intermediate point to transfer relative phases.

8. From *(Fix absolute amplitude for $\rho_{ij}^{(\mathrm{triplet,i})}$)*, *(Transfer relative phase $+1$, a")*, and *(Transfer relative phase $+\mathrm{i}$, b")*, we have

$$(U^{(2)})^{-1}\rho_{ij}^{(\mathrm{triplet,i})}(U^{(2)}) = \frac{1}{3}\left(|1\rangle + |i\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle 1| + \langle i| - \mathrm{i}\,\langle j|\right), \quad \forall i \neq j \in \{2,\ldots,d\}. \tag{D90}$$

9. From *(Transfer relative phase $+\mathrm{i}$, c")*, we have

$$(U^{(2)})^{-1}\rho_{ij}^{(\mathrm{imag})}(U^{(2)}) = \frac{1}{2}\left(|i\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle i| - \mathrm{i}\,\langle j|\right), \quad \forall i \neq j \in \{2,\ldots,d\}. \tag{D91}$$

Here, the state $\rho_{ij}^{(\mathrm{triplet,i})}$ is used as an intermediate point to transfer relative phases.

By considering $U = U^{(2)}$, we have established the claim. □

As η goes to zero, the learning algorithm can find a set of states that satisfy the geometry up to an arbitrarily small error. This implies that the description for the states would also be close to the true one up to an arbitrarily small error as η goes to zero. Using this basic, idea, we can see that Lemma D.11 yields the corollary stated below.

Corollary D.12. *Given $\epsilon > 0$. For all sufficiently small η, there exists a unitary or anti-unitary transformation U, such that*

$$\left\|\rho_i^{(\mathrm{basis})} - U|i\rangle\langle i|U^{-1}\right\| < \epsilon, \qquad \forall i \in \{1,\ldots,d\}, \tag{D92}$$

$$\left\|\rho_{ij}^{(\mathrm{real})} - \frac{1}{2}U\left(|i\rangle + |j\rangle\right)\left(\langle i| + \langle j|\right)U^{-1}\right\| < \epsilon, \qquad \forall i \neq j \in \{1,\ldots,d\}, \tag{D93}$$

$$\left\|\rho_{ij}^{(\mathrm{imag})} - \frac{1}{2}U\left(|i\rangle + \mathrm{i}\,|j\rangle\right)\left(\langle i| - \mathrm{i}\,\langle j|\right)U^{-1}\right\| < \epsilon, \qquad \forall i \neq j \in \{2,\ldots,d\}. \tag{D94}$$

We only need to focus on these three sets of states in the following discussion, but the claim also holds for the other sets of states.

D.6. Quantum state/process/measurement tomography

In the final step, the learning algorithm utilizes the learned descriptions of the states in the previous subsection to perform quantum state/process/measurement tomography. For readers familiar with quantum tomography, the claim could be established easily. For completeness, we present the detailed derivations in the following.

a. States: $\forall x \in \mathcal{X}$, we can always write ρ_x as

$$\rho_x = U \sum_{ij} a_{ij} |i\rangle\langle j| U^{-1}, \tag{D95}$$

where $(a_{ij})_{ij}$ is a Hermitian matrix. If we write

$$\sum_{ij} a_{ij} |i\rangle\langle j| = \sum_i a_{ii} |i\rangle\langle i| + \sum_{i\neq j} \frac{a_{ij}+a_{ji}}{2} (|i\rangle\langle j| + |j\rangle\langle i|) + \sum_{i\neq j} \frac{a_{ij}-a_{ji}}{2} (|i\rangle\langle j| - |j\rangle\langle i|), \tag{D96}$$

and we assume that Eq. (D92), (D93), and (D94) holds exactly, then we can learn the matrix $(a_{ij})_{ij}$ by noting the following identities

$$a_{ii} = \mathrm{tr}(\rho_x \rho_i^{(\mathrm{basis})}), \ \forall i, \tag{D97}$$

$$\frac{a_{ij}+a_{ji}}{2} = \mathrm{tr}(\rho_x \rho_{ij}^{(\mathrm{real})}) - \frac{\mathrm{tr}(\rho_x \rho_i^{(\mathrm{basis})}) + \mathrm{tr}(\rho_x \rho_j^{(\mathrm{basis})})}{2}, \ \forall i \neq j, \tag{D98}$$

$$\frac{a_{ij}-a_{ji}}{2} = \frac{1}{\mathrm{i}} \left[\mathrm{tr}(\rho_x \rho_{ij}^{(\mathrm{imag})}) - \frac{\mathrm{tr}(\rho_x \rho_i^{(\mathrm{basis})}) + \mathrm{tr}(\rho_x \rho_j^{(\mathrm{basis})})}{2} \right], \ \forall i \neq j. \tag{D99}$$

The state overlap $\mathrm{tr}(\rho_1\rho_2)$ can be estimated using the procedure provided in Appendix D.4 based on approximate Haar random unitaries. For $\eta > 0$, define $\tilde{\rho}_x$ to be the empirical estimate of $\sum_{ij} a_{ij}|i\rangle\langle j|$ based on the above equations. Due to error in the estimation of the state overlap $\mathrm{tr}(\rho_1\rho_2)$ and the error in the states $\rho_i^{(\mathrm{basis})}, \rho_{ij}^{(\mathrm{real})}, \rho_{ij}^{(\mathrm{imag})}$, we have $\tilde{\rho}_x \neq \sum_{ij} a_{ij}|i\rangle\langle j|$. Nevertheless, one can use basic inequalities to show that $\left\| \tilde{\rho}_x - \sum_{ij} a_{ij}|i\rangle\langle j| \right\|_1 < e_a(\eta)$ with high probability. The error $e_a(\eta)$ can be made arbitrarily small when η goes to zero.

b. POVMs: $\forall z \in \mathcal{Z}, \forall b \in \mathcal{B}$, we can learn M_{zb} similar to learning states. We write M_{zb} as

$$M_{zb} = U \sum_{ij} b_{ij} |i\rangle\langle j| U^{-1}, \tag{D100}$$

where $(b_{ij})_{ij}$ is a Hermitian matrix. For any quantum state ρ, we can estimate $\mathrm{tr}(M_{zb}\rho)$ by simply computing the proportion of counts that we see the outcome b when we measure $\mathcal{M}_z$ on ρ. Using this simple procedure, we can estimate $\mathrm{tr}(M_{zb}\rho)$ to an error η with high probability. Then we learn the matrix $(b_{ij})_{ij}$ using the same formulas given in Eq. (D97) (D98) and (D99), but we replace ρ_x with M_{zb}. For $\eta > 0$, we define $\tilde{M}_{zb}$ to be the empirical estimate of $\sum_{ij} b_{ij}|i\rangle\langle j|$. Due to error in the estimation of $\mathrm{tr}(M_{zb}\rho)$ and the error in $\rho_i^{(\mathrm{basis})}, \rho_{ij}^{(\mathrm{real})}, \rho_{ij}^{(\mathrm{imag})}$, $\tilde{M}_{zb}$ is not exactly equal to $\sum_{ij} b_{ij}|i\rangle\langle j|$. But $\tilde{M}_{zb}$ will be close to $\sum_{ij} b_{ij}|i\rangle\langle j|$. In particular, there exists an error function $e_b(\eta)$, such that $\lim_{\eta\to 0} e_b(\eta) = 0$ and $\left\| \tilde{M}_{zb} - \sum_{ij} b_{ij}|i\rangle\langle j| \right\|_1 < e_b(\eta)$ with high probability for any η.

c. CPTP maps: $\forall y \in \mathcal{Y}$, we can write $\mathcal{E}_y$ as

$$\mathcal{E}_y(\cdot) = U \sum_{ijkl} c_{ijkl} |k\rangle\langle l| \, \mathrm{tr}(|i\rangle\langle j| U^{-1}(\cdot)U) U^{-1}. \tag{D101}$$

The coefficients c_{ijkl} could be learned using the state overlap procedure given in Appendix D.4 and the states $\rho_i^{(\mathrm{basis})}, \rho_{ij}^{(\mathrm{real})}, \rho_{ij}^{(\mathrm{imag})}$ in Eq. (D92), (D93), and (D94). To achieve this, we gather a collection

of data by preparing each state in $\rho_i^{(\text{basis})}, \rho_{ij}^{(\text{real})}, \rho_{ij}^{(\text{imag})}$, evolving under $\mathcal{E}_y$, and estimating the state overlap of the output state with every state in $\rho_i^{(\text{basis})}, \rho_{ij}^{(\text{real})}, \rho_{ij}^{(\text{imag})}$. We can then use the collection of data to estimate $c_{ijkl}, \forall ijkl$. If Eq. (D92), (D93), and (D94) holds exactly, then for $i \neq j$ and $k = l$,

$$c_{ijkl} = \text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_{ij}^{(\text{real})})) + \frac{1}{\text{i}} \text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_{ij}^{(\text{imag})})) \tag{D102}$$

$$- \left(\frac{1}{2} + \frac{1}{2\text{i}}\right) \left(\text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_i^{(\text{basis})})) + \text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_j^{(\text{basis})}))\right) \tag{D103}$$

For $i = j$ and $k = l$, we can obtain

$$c_{ijkl} = \text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_i^{(\text{basis})})). \tag{D104}$$

For $i = j$ and $k \neq l$, we see that

$$c_{ijkl} = \text{tr}(\rho_{kl}^{(\text{real})} \mathcal{E}_y(\rho_i^{(\text{basis})})) + \frac{1}{\text{i}} \text{tr}(\rho_{kl}^{(\text{imag})} \mathcal{E}_y(\rho_i^{(\text{basis})})) \tag{D105}$$

$$- \left(\frac{1}{2} + \frac{1}{2\text{i}}\right) \left(\text{tr}(\rho_k^{(\text{basis})} \mathcal{E}_y(\rho_i^{(\text{basis})})) + \text{tr}(\rho_l^{(\text{basis})} \mathcal{E}_y(\rho_i^{(\text{basis})}))\right) \tag{D106}$$

For $i \neq j$ and $k \neq l$, we have

$$c_{ijkl} = \text{tr}\left(\left(\rho_{kl}^{(\text{real})} + \frac{1}{\text{i}}\rho_{kl}^{(\text{imag})} - \left(\frac{1}{2} + \frac{1}{2\text{i}}\right)\left(\rho_k^{(\text{basis})} + \rho_l^{(\text{basis})}\right)\right)\right. \tag{D107}$$

$$\left.\mathcal{E}_y\left(\left(\rho_{ij}^{(\text{real})} + \frac{1}{\text{i}}\rho_{ij}^{(\text{imag})} - \left(\frac{1}{2} + \frac{1}{2\text{i}}\right)\left(\rho_i^{(\text{basis})} + \rho_j^{(\text{basis})}\right)\right)\right)\right). \tag{D108}$$

Expanding the right hand side of the above equation gives a weighted sum of $\text{tr}(\rho_2 \mathcal{E}_y(\rho_1))$ for some states ρ_1, ρ_2. We consider $\tilde{\mathcal{E}}_y(\cdot)$ to be the empirical estimate for $\sum_{ijkl} c_{ijkl} |k\rangle\langle l| \,\text{tr}(|i\rangle\langle j|(\cdot))$. There exists an error function $e_c(\eta)$ such that $\left\|\tilde{\mathcal{E}}_y(\cdot) - \sum_{ijkl} c_{ijkl} |k\rangle\langle l| \,\text{tr}(|i\rangle\langle j|(\cdot))\right\|_\diamond < e_c(\eta)$. Furthermore, as η approaches zero, $e_c(\eta)$ goes to zero.

D.7. Putting everything together

For all world model $\mathcal{W} \in \mathcal{Q}$, after finishing the tomography step in Appendix D.6, we can guarantee the following. There is an error function $\epsilon(\eta)$. For $\eta > 0$, there exists a global unitary or anti-unitary transformation U, such that the learned descriptions $\tilde{\rho}_x, \tilde{\mathcal{E}}_y, \tilde{M}_{zb}$ satisfies

$$\left\|\rho_x - U\tilde{\rho}_x U^{-1}\right\|_1 < \epsilon(\eta), \tag{D109}$$

$$\left\|M_{zb} - U\tilde{M}_{zb}U^{-1}\right\|_1 < \epsilon(\eta), \tag{D110}$$

$$\left\|\mathcal{E}_y(\cdot) - U\tilde{\mathcal{E}}_y(U^{-1}(\cdot)U)U^{-1}\right\|_\diamond < \epsilon(\eta), \tag{D111}$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$. As η goes to zero, $\epsilon(\eta)$ goes to zero. After running the above procedures with precision parameter η, the learning algorithm considers $\eta \leftarrow \eta/2$ and repeatedly runs the previous steps to obtain more accurate descriptions. Because $\epsilon(\eta)$ goes to zero as η goes to zero, the learning algorithm can learn the all the physical descriptions to arbitrarily small error up to a global unitary or anti-unitary transformation. Hence, $\mathcal{Q}$ is learnable. This concludes the proof of Theorem D.1.

E. Basic properties of learnability and unlearnability

In this section, we will present various basic results regarding the relationship between different classes of world models. These results will be useful for proving what kinds of world models are learnable, and what kinds are not, in the following sections.

We begin with a basic property: If a model class $\mathcal{Q}$ is learnable, then any subset $\mathcal{Q}'$ of $\mathcal{Q}$ is also learnable. This property is expected because removing possible models from the class will not make it harder to learn which model is the correct one. We note that $\mathcal{Q}'$ and $\mathcal{Q}$ are model classes, i.e., sets of potential world models. The relation of $\mathcal{Q}' \subseteq \mathcal{Q}$ is very different from the concept of extension given in Definition A.4, which considers relation between two world models.

Proposition E.1 (Monotonicity of (un)learnability)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and two model classes $\mathcal{Q}, \mathcal{Q}'$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ such that $\mathcal{Q}' \subseteq \mathcal{Q}$. If $\mathcal{Q}$ is learnable, then $\mathcal{Q}'$ is learnable. Equivalently, if $\mathcal{Q}'$ is unlearnable, then $\mathcal{Q}$ is unlearnable.*

Proof. Every world model in $\mathcal{Q}'$ is in $\mathcal{Q}$. Hence, if $\mathcal{Q}$ is learnable, then $\mathcal{Q}' \subseteq \mathcal{Q}$ is learnable. This is equivalent to the contrapositive statement: if $\mathcal{Q}' \subseteq \mathcal{Q}$ is unlearnable, then $\mathcal{Q}$ is unlearnable. □

Another important result states that a model class $\mathcal{Q}$ is unlearnable if $\mathcal{Q}$ contains world models $\mathcal{W}_1$ and $\mathcal{W}_2$ that are weakly indistinguishable, but are not equivalent. This follows because, by the definition of weakly indistinguishable, no experiment within the model class can tell $\mathcal{W}_1$ and $\mathcal{W}_2$ apart. It may seem that this follows immediately from the definition, but there are subtlety arising from the fact that learning is probabilistic and allows arbitrarily small error.

Proposition E.2 (Weakly indistinguishability implies unlearnability)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ for d-dimensional world models over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. If there exists $\mathcal{W}_1 \not\equiv \mathcal{W}_2 \in \mathcal{Q}$ such that $\mathcal{W}_1$ and $\mathcal{W}_2$ are weakly indistinguishable, then $\mathcal{Q}$ is unlearnable.*

Proof. Assume that $\mathcal{W}_A \not\equiv \mathcal{W}_B \in \mathcal{Q}$ are weakly indistinguishable. Because $\mathcal{W}_A \not\equiv \mathcal{W}_B$, for all unitary and anti-unitary U, there exists $x \in \mathcal{X}$ or $y \in \mathcal{Y}$ or $z \in \mathcal{Z}, b \in \mathcal{B}$ such that the corresponding physical operations are different, i.e., $\rho_x^B \neq U\rho_x^A U^{-1}$, $\mathcal{E}_y^B(\cdot) \neq U\mathcal{E}_y^A(U^{-1}(\cdot)U)U^{-1}$, or $M_{zb}^B \neq UM_{zb}^A U^{-1}$. We define the minimum error $\tilde{\epsilon}$ over U to be

$$\tilde{\epsilon} = \min_U \sup_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}\\ z\in\mathcal{Z}\\ b\in\mathcal{B}}} \left(\left\|\rho_x^B - U\rho_x^A U^{-1}\right\|_1, \left\|\mathcal{E}_y^B(\cdot) - U\mathcal{E}_y^A(U^{-1}(\cdot)U)U^{-1}\right\|_\diamond, \left\|M_{zb}^B - UM_{zb}^A U^{-1}\right\|_1 \right), \tag{E1}$$

where U is a unitary or anti-unitary transformation. We can use minimum instead of infimum because unitary and anti-unitary transformations form a compact space. If $\tilde{\epsilon} = 0$, then $\mathcal{W}_A \equiv \mathcal{W}_B$. Hence $\tilde{\epsilon} > 0$. The quantity $\tilde{\epsilon}$ sets an lower bound on the error for what any algorithm could learn.

Suppose that $\mathcal{Q}$ is learnable. Then, there is an algorithm $\mathcal{A}$, for $\delta = 1/3$ and $\epsilon = \tilde{\epsilon}/3$, for world model $\mathcal{W}_A$, there exists a unitary or anti-unitary U_A, such that for any action, with probability at least $1-\delta$, the output from the algorithm has an error of at most ϵ after transforming under U_A. Similarly, for world model $\mathcal{W}_B$, there exists a unitary or anti-unitary U_B, such that for any action, with probability at least $1-\delta$, the output from the algorithm has an error of at most ϵ after transforming under U_B. Consider $U_* = U_B U_A^{-1}$. From the definition of $\tilde{\epsilon}$, we have

$$\sup_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}\\ z\in\mathcal{Z}\\ b\in\mathcal{B}}} \left(\left\|\rho_x^B - U_*\rho_x^A U_*^{-1}\right\|_1, \left\|\mathcal{E}_y^B(\cdot) - U_*\mathcal{E}_y^A(U_*^{-1}(\cdot)U_*)U_*^{-1}\right\|_\diamond, \left\|M_{zb}^B - U_* M_{zb}^A U_*^{-1}\right\|_1 \right) \geq \tilde{\epsilon}. \tag{E2}$$

Hence, there exists an action such that the error $> \frac{9}{10}\tilde{\epsilon}$. Without loss of generality, assume that for some $x \in \mathcal{X}$, $\left\|\rho_x^B - U_*\rho_x^A U_*^{-1}\right\|_1 > \frac{9}{10}\tilde{\epsilon}$.

Because $\mathcal{W}_A, \mathcal{W}_B$ are weakly indistinguishable, the output $\tilde{\rho}_x$ will have the same probability distribution for world model $\mathcal{W}_A$ and $\mathcal{W}_B$. This is an immediate consequence of Eq. (A6) in Definition A.6. We will refer to the probability distribution as $p_{AB}(\tilde{\rho}_x)$. From the learnability, we have

$\left\|\rho_x^A - U_A\tilde{\rho}_x U_A^{-1}\right\|_1 \leq \epsilon$ with probability $\geq 2/3$ and $\left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 \leq \epsilon$ with probability $\geq 2/3$. The above probability statement are over the same probability distribution $p_{AB}(\tilde{\rho}_x)$. Equivalently,

$$P\left[\left\|\rho_x^A - U_A\tilde{\rho}_x U_A^{-1}\right\|_1 > \epsilon\right] < 1/3, \tag{E3}$$

$$P\left[\left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 > \epsilon\right] < 1/3, \tag{E4}$$

Hence, by union bound, we have

$$P\left[\left\|\rho_x^A - U_A\tilde{\rho}_x U_A^{-1}\right\|_1 > \epsilon \text{ or } \left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 > \epsilon\right] < 2/3. \tag{E5}$$

This is equivalent to the fact that $\left\|\rho_x^A - U_A\tilde{\rho}_x U_A^{-1}\right\|_1 \leq \epsilon$ and $\left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 \leq \epsilon$ with probability $\geq 1/3$. Because the probability is greater than zero and the probability distribution is over the choice of $\tilde{\rho}_x$, there exists $\tilde{\rho}_x \in \mathbb{C}^{d\times d}$ such that $\left\|\rho_x^A - U_A\tilde{\rho}_x U_A^{-1}\right\|_1 \leq \epsilon$ and $\left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 \leq \epsilon$. Because trace norm $\|\cdot\|_1$ is invariant under unitary or anti-unitary transformation, we have $\left\|U_B U_A^{-1}\rho_x^A U_A U_B^{-1} - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 \leq \epsilon$. Recall that $U_* = U_B U_A^{-1}$. By triangle inequality, we have

$$\frac{9}{10}\tilde{\epsilon} < \left\|U_*\rho_x^A U_*^{-1} - \rho_x^B\right\|_1 \leq \left\|U_B U_A^{-1}\rho_x^A U_A U_B^{-1} - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 + \left\|\rho_x^B - U_B\tilde{\rho}_x U_B^{-1}\right\|_1 \leq 2\epsilon = \frac{2}{3}\tilde{\epsilon}. \tag{E6}$$

This is a contradiction, hence $\mathcal{Q}$ is not learnable. □

We present a simple example where $\mathcal{W}_1$ and $\mathcal{W}_2$ are weakly indistinguishable, but not equivalent. Consider a model class $\mathcal{Q}$ that contains two $d = 2$-dimensional world models $\{\mathcal{W}_A, \mathcal{W}_B\}$ with the action space $\mathcal{X} = \{0\}, \mathcal{Y} = \{h, t\}, \mathcal{Z} = \{0\}$ and the outcome space $\mathcal{B} = \{0, 1\}$. We define the physical actions in $\mathcal{W}_A$ and $\mathcal{W}_B$ as

$$\rho_0^A = I/2, \qquad \mathcal{E}_h^A(\rho) = H\rho H^\dagger, \qquad \mathcal{E}_t^A(\rho) = T\rho T^\dagger, \qquad \mathcal{M}_0^A = \{I/2, I/2\}, \tag{E7}$$

$$\rho_0^B = |0\rangle\langle 0|, \qquad \mathcal{E}_h^B(\rho) = H\rho H^\dagger, \qquad \mathcal{E}_t^B(\rho) = T\rho T^\dagger, \qquad \mathcal{M}_0^B = \{I/2, I/2\}. \tag{E8}$$

$\mathcal{W}_A$ has an initial state that is maximally mixed, hence the state ρ_0^A has a purity $\mathrm{tr}((\rho_0^A)^2)$ of $1/2$. But $\mathcal{W}_B$ has an initial state that is pure, so the state ρ_0^B has a purity of 1. Theorem A.8 implies that the two world models are not equivalent. However, both of the POVMS $\mathcal{M}_0^A$ and $\mathcal{M}_0^B$ produce uniformly random outcomes in $\mathcal{B}$ when applied to any state. Therefore, $\mathcal{W}_A$ and $\mathcal{W}_B$ are weakly indistinguishable, and hence by Proposition E.2 the model class $\mathcal{Q}$ is unlearnable. In this example, both of the world models $\mathcal{W}_A$ and $\mathcal{W}_B$ have a useless measurement device that provides no information, so there is no way to learn which is which.

Monotonicity of learnability focuses on two model classes that have the same action spaces. Here, we provide a basic proposition that considers two model classes with different action spaces. The proposition holds because of the compositional nature in the design of an experiment — we can compose different states, evolutions, and POVMs to form new states, evolutions, and POVMs.

Proposition E.3 (Learnability after adding composed states)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Consider a set Ξ, a constant $L \geq 1$, and a function f that takes in $\xi \in \Xi$ and outputs $(x, y_1, \ldots, y_\ell)$ where $\ell \leq L, x \in \mathcal{X}, y_1, \ldots, y_\ell \in \mathcal{Y}$. The model class $\mathcal{Q}'$ is over $\mathcal{X}' = \mathcal{X} \cup \Xi, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, and contains the world model*

$$\left(\{\rho_x\}_{x\in\mathcal{X}} \cup \{\rho_\xi = (\mathcal{E}_{y_\ell} \circ \ldots \circ \mathcal{E}_{y_1})(\rho_x)\}_{\substack{\xi\in\Xi, \\ f(\xi)=(x,y_1,\ldots,y_\ell)}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}}\right), \tag{E9}$$

for each world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ in the original model class $\mathcal{Q}$. We have $\mathcal{Q}$ is learnable if and only if $\mathcal{Q}'$ is learnable.

Proof. We begin with two basic statements. First, every experiment in $\mathcal{Q}$ can be simulated by an experiment in $\mathcal{Q}'$. And every experiment in $\mathcal{Q}'$ can be simulated by an experiment in $\mathcal{Q}$. The first statement immediately holds by noting that $\mathcal{Q}'$ contains all the actions in $\mathcal{Q}$. The second statement is true because the new action added in $\mathcal{Q}'$ is composed of actions in $\mathcal{Q}$. Since each experiment E is a composition of actions, we can compose the corresponding actions in $\mathcal{Q}$ to simulate an experiment in $\mathcal{Q}'$. The proof of this proposition is simple given the knowledge of these two facts. We separate the proof for the two directions of the statement into two paragraphs.

a. $\mathcal{Q}'$ is learnable implies $\mathcal{Q}$ is learnable: If $\mathcal{Q}'$ is learnable, then there is a learning algorithm, such that for every action a in the action spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, the algorithm uses actions in $\mathcal{X}', \mathcal{Y}, \mathcal{Z}$ to learn the intrinsic description of action a. We can simulate every added action ξ in Ξ with the actions $x \in \mathcal{X}, y_1, \ldots, y_\ell \in \mathcal{Y}$, where $f(\xi) = (x, y_1, \ldots, y_\ell)$. Hence, we have a learning algorithm using only actions in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ to learn the description of actions in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ under the model class $\mathcal{Q}$. Together, $\mathcal{Q}$ is learnable if $\mathcal{Q}'$ is learnable.

b. $\mathcal{Q}$ is learnable implies $\mathcal{Q}'$ is learnable: If $\mathcal{Q}$ is learnable, then there is a learning algorithm that learns the physical operation associated with every actions in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. To show that $\mathcal{Q}'$ is learnable, we need to show that all actions in $\mathcal{X}', \mathcal{Y}, \mathcal{Z}$ are learnable. By simulating the experiments under $\mathcal{Q}$ using actions in $\mathcal{Q}'$, we can learn all the physical operations associated to actions in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. Now for all $\xi \in \Xi$, the initial state ρ_ξ associated to the additional action ξ with $f(\xi) = (x, y_1, \ldots, y_\ell)$ can be learned. This follows from the facts that ρ_ξ is equal to $(\mathcal{E}_{y_\ell} \circ \ldots \circ \mathcal{E}_{y_1})(\rho_x)$, and each of $\rho_x, \mathcal{E}_{y_1}, \ldots, \mathcal{E}_{y_\ell}$ can be learned to arbitrarily high accuracy up to a global unitary or anti-unitary transformation. Therefore, $\mathcal{Q}'$ is learnable. □

The same proof for Proposition E.3 can be used to establish the following other two propositions where we consider a model class with new composed CPTP maps or POVM.

Proposition E.4 (Learnability after adding composed CPTP maps)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Consider a set Ξ, a constant $L \geq 2$, and a function f that takes in an element ξ in Ξ and outputs $(y_1, \ldots, y_\ell)$ where $\ell \leq L, y_1, \ldots, y_\ell \in \mathcal{Y}$. The model class $\mathcal{Q}'$ is over $\mathcal{X}, \mathcal{Y}' = \mathcal{Y} \cup \Xi, \mathcal{Z}, \mathcal{B}$, and contains the world model*

$$\left(\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}} \cup \{\mathcal{E}_\xi = (\mathcal{E}_{y_\ell} \circ \ldots \circ \mathcal{E}_{y_1})\}_{\substack{\xi\in\Xi, \\ f(\xi)=(y_1,\ldots,y_\ell)}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}} \right), \tag{E10}$$

for each world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ in the original model class $\mathcal{Q}$. We have $\mathcal{Q}$ is learnable if and only if $\mathcal{Q}'$ is learnable.

Proposition E.5 (Learnability after adding composed POVMs)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Consider a set Ξ, a constant $L \geq 1$, and a function f that takes in an element ξ in Ξ and outputs $(y_1, \ldots, y_\ell, z)$ where $\ell \leq L, y_1, \ldots, y_\ell \in \mathcal{Y}, z \in \mathcal{Z}$. The model class $\mathcal{Q}'$ is over the spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}' = \mathcal{Z} \cup \Xi, \mathcal{B}$, and contains the world model*

$$\left(\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}} \cup \{\mathcal{M}_\xi = \mathcal{M}_z \circ (\mathcal{E}_{y_\ell} \circ \ldots \circ \mathcal{E}_{y_1})\}_{\substack{\xi\in\Xi, \\ f(\xi)=(y_1,\ldots,y_\ell,z)}} \right), \tag{E11}$$

for each world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ in the original model class $\mathcal{Q}$. We have $\mathcal{Q}$ is learnable if and only if $\mathcal{Q}'$ is learnable.

Similar to the three propositions stated above, we can consider new actions that are convex combinations of existing actions. Adding the new actions does not affect the learnability.

Proposition E.6 (Learnability after adding mixtures of states)**.** *Given sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ and a model class $\mathcal{Q} = \{\mathcal{W}\}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$. Consider a set Ξ, a constant $L \geq 1$, and a function f that takes in an element ξ in Ξ and outputs $((p_1, x_1) \ldots, (p_\ell, x_\ell))$ where $\ell \leq L, x_1, \ldots, x_\ell \in \mathcal{X}$ and $(p_1, \ldots, p_\ell)$ is a probability distribution. The model class $\mathcal{Q}'$ is over the spaces $\mathcal{X}' = \mathcal{X} \cup \Xi, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, and is contains the world model*

$$\left(\{\rho_x\}_{x\in\mathcal{X}} \cup \left\{ \rho_\xi = \sum_{\ell=1}^{L} p_\ell \rho_{x_\ell} \right\}_{\substack{\xi\in\Xi, \\ f(\xi)=((p_1,x_1)\ldots,(p_\ell,x_\ell))}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}} \right), \tag{E12}$$

for each world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ in the original model class $\mathcal{Q}$. We have $\mathcal{Q}$ is learnable if and only if $\mathcal{Q}'$ is learnable.

Proof. Similar to the proof of Proposition E.3, every experiment in $\mathcal{Q}$ can be simulated by experiments in $\mathcal{Q}'$. And every experiment in $\mathcal{Q}'$ can be simulated by experiments in $\mathcal{Q}$. The first statement is trivial as $\mathcal{Q}'$ contains all actions in $\mathcal{Q}$. The second statement is true because we can simulate any experiment that begins with the action $\xi \in \Xi$ by randomly sampling x_i from $(x_1, \ldots, x_\ell)$ according to the probability distribution $(p_1, \ldots, p_\ell)$ and running the experiment using $x_i \in \mathcal{X}$. Using essential the same proof as Proposition E.3, we can show that $\mathcal{Q}'$ is learnable implies $\mathcal{Q}$ is learnable and $\mathcal{Q}$ is learnable implies $\mathcal{Q}'$ is learnable. □

Using essentially the same proof as Proposition E.6, we can also obtain the equivalence of learnability after adding mixtures of CPTP maps or POVMs.

F. Gate-dependent Pauli noises is unlearnable with Clifford+T gates

F.1. Statement and unlearnability of a target model class

In quantum state/process tomography, it is well known that Clifford circuits are informationally complete. For example, we can learn any quantum state with Clifford circuits and computational basis measurement. We can also learn any quantum process with Clifford circuits, all zero state preparation, and computational basis measurements. In these works, it is often assumed that the Clifford circuits, the state preparation, and the measurements are perfect. The situation changes dramatically when these physical operations are not perfect.

In this section, we show that when there are gate-dependent Pauli noise, Clifford circuits are fundamentally uncapable of learning the noise processes. Even more interestingly, adding T gate is still insufficient.

Theorem F.1 (Restatement of Theorem 6; Gate-dependent Pauli noise is unlearnable with Clifford+T gates)**.** *Given $\frac{1}{2} > \epsilon > 0$. Consider a qubit system. Suppose we can prepare the zero state $|0\rangle$ perfectly and any state ρ with an unknown error $\leq \epsilon$, measure in the computational basis perfectly, and apply Clifford gates and T gate, where each gate is followed by an unknown gate-dependent Pauli noise channel that is ϵ-close to the identity channel. It is impossible for any algorithm to learn the gate-dependent Pauli noise channels to arbitrarily small error.*

To prove Theorem F.1, we begin by stating the conditions in Theorem F.1 as a model class. Consider the the action spaces $\mathcal{X} = \{x_\sigma\}_{\sigma:\text{state}}, \mathcal{Y} = \{y_U\}_{U\in\mathcal{C}\cup\{T\}}, \mathcal{Z} = \{0\}$, where σ is a quantum state, $\mathcal{C}$ is the Clifford group, and T is the T gate. And consider the outcome space $\mathcal{B} = \{0, 1\}$. Given $\epsilon > 0$, we define the model class $\mathcal{Q}^\epsilon = \{\mathcal{W}\}$ to be the set of world models $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ that satisfies the following conditions

$$\|\rho_{x_\sigma} - \sigma\|_1 \leq \epsilon, \qquad \forall\sigma : \text{state}, \tag{F1}$$

$$\rho_{x_{|0\rangle\langle 0|}} = |0\rangle\langle 0|, \tag{F2}$$

$$\mathcal{E}_{y_U}(\rho) = \mathcal{P}_U(U\rho U^\dagger),\ \|\mathcal{P}_U - \mathcal{I}\|_\diamond \leq \epsilon, \qquad \forall U \in \mathcal{C}\cup\{T\}, \tag{F3}$$

$$\mathcal{M}_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}. \tag{F4}$$

where $\mathcal{P}_U$ is a Pauli channel, i.e.,

$$\mathcal{P}_U(\rho) = p_I^U\rho + p_X^U X\rho X + p_Y^U Y\rho Y + p_Z^U Z\rho Z \tag{F5}$$

for some probability distribution $\left(p_I^U, p_X^U, p_Y^U, p_Z^U\right)$, or equivalently

$$\mathcal{P}_U(I) = I, \mathcal{P}_U(X) = \lambda_X^U X, \mathcal{P}_U(Y) = \lambda_Y^U Y, \mathcal{P}_U(Z) = \lambda_Z^U Z \tag{F6}$$

for some real value $\lambda_X^U, \lambda_Y^U, \lambda_Z^U$, $\mathcal{I}$ is the identity channel, and $\|\cdot\|_\diamond$ is the diamond norm.

F.2. Unlearnability of $\mathcal{Q}^\epsilon$

We provide more analysis of $\mathcal{Q}^\epsilon$ in this subsection. First, we show that $\mathcal{Q}^\epsilon$ is not redundant: no distinct world models in $\mathcal{Q}^\epsilon$ are equivalent and hence describe the same physical reality. From Theorem A.8, two equivalent world models are related by a global unitary or anti-unitary transformation U. If two distinct world models in $\mathcal{Q}^\epsilon$ are equivalent, then there exists a global unitary or anti-unitary transformation $U \neq \mathrm{e}^{\mathrm{i}\phi} I$, for some $\phi \in \mathbb{R}$, and Pauli channels $\mathcal{P}, \mathcal{P}', \mathcal{Q}, \mathcal{Q}'$ that are at most ϵ-far from identity channel, such that

$$|0\rangle\langle 0| = U|0\rangle\langle 0|U^{-1}, \tag{F7}$$

$$|1\rangle\langle 1| = U|1\rangle\langle 1|U^{-1}, \tag{F8}$$

$$\mathcal{P}(H|0\rangle\langle 0|H) = U\mathcal{P}'(HU^{-1}|0\rangle\langle 0|UH)U^{-1}, \tag{F9}$$

$$\mathcal{Q}((SH)|0\rangle\langle 0|(SH)^\dagger) = U\mathcal{Q}'((SH)U^{-1}|0\rangle\langle 0|U(SH)^\dagger)U^{-1}, \tag{F10}$$

where H is the Hadamard gate and S is the phase gate. These conditions imply that

$$Z = UZU^{-1}, \lambda X = \lambda' UXU^{-1}, \Lambda Y = \Lambda' UYU^{-1}, \tag{F11}$$

where $\lambda, \lambda', \Lambda, \Lambda' \geq 1/2$ from Eq. (F12), $|+\rangle, |y+\rangle$ are the eigenvector with eigenvalue $+1$ for X, Y. Because a unitary $U = \mathrm{e}^{\mathrm{i}\phi}\exp(\mathrm{i}(aX+bY+cZ))$ and an anti-unitary $U = \mathrm{e}^{\mathrm{i}\phi}\exp(\mathrm{i}(aX+bY+cZ))K$ for $a, b, c, \phi \in \mathbb{R}$ and K the complex conjugation, Eq. (F11) implies that $U = \mathrm{e}^{\mathrm{i}\phi} I$, which is a contradiction. Therefore, no distinct world models in $\mathcal{Q}^\epsilon$ are equivalent.

The fact that $\mathcal{Q}^\epsilon$ is not redundant is useful for the following logical reasoning. Suppose no algorithm can learn the gate-dependent Pauli noise channels to arbitrarily small error. Then we can find two distinct world models in $\mathcal{Q}$ such that the two world models are weakly indistinguishable. From the non-redundancy of $\mathcal{Q}^\epsilon$, these two distinct world models are not equivalent to one another. Hence from Proposition E.2, the model class $\mathcal{Q}^\epsilon$ is unlearnable.

On the other hand, suppose there is an algorithm that can learn the gate-dependent Pauli noise channels to arbitrarily small error. Recall that for two Pauli channels $\mathcal{P}, \mathcal{P}'$, $\|\mathcal{P} - \mathcal{P}'\|_\diamond = |p_I - p'_I| + |p_X - p'_X| + |p_Y - p'_Y| + |p_Z - p'_Z|$ and $\lambda_X = 1 - 2p_Y - 2p_Z, \lambda_Y = 1 - 2p_X - 2p_Z, \lambda_Z = 1 - 2p_X - 2p_Y$. Therefore, $\|\mathcal{P} - \mathcal{I}\| < \epsilon \leq 1/2$ implies

$$2(p_X + p_Y + p_Z) \leq 1/2,\ \lambda_X, \lambda_Y, \lambda_Z \geq 1/2. \tag{F12}$$

We can build on the Pauli channel learning algorithm to learn ρ_{x_σ} for any state σ using the following equation

$$\rho_{x_\sigma} = \frac{I}{2} + \mathrm{tr}(X\rho_{x_\sigma})\frac{X}{2} + \mathrm{tr}(Y\rho_{x_\sigma})\frac{Y}{2} + \mathrm{tr}(Z\rho_{x_\sigma})\frac{Z}{2} \tag{F13}$$

$$= \frac{I}{2} + \frac{\mathrm{tr}(Z\mathcal{E}_{y_H}(\rho_{x_\sigma}))}{\lambda_Z^H}\frac{X}{2} - \frac{\mathrm{tr}(Z\mathcal{E}_{y_S}(\mathcal{E}_{y_H}(\rho_{x_\sigma})))}{\lambda_Z^H\lambda_X^S}\frac{Y}{2} + \mathrm{tr}(Z\rho_{x_\sigma})\frac{Z}{2}, \tag{F14}$$

and $\mathrm{tr}(Z\rho) = \langle 0|\rho|0\rangle - \langle 1|\rho|1\rangle$. Because every action in $\mathcal{Q}^\epsilon$ can be learned to arbitrarily high accuracy, the model class $\mathcal{Q}^\epsilon$ is learnable. Together, Theorem F.1 is equivalent to stating that $\mathcal{Q}^\epsilon$ is unlearnable. In the following, we will prove that $\mathcal{Q}^\epsilon$ is unlearnable.

F.3. Unlearnability of a simpler model class

We combine the basic results proven in Appendix E to establish the unlearnability of $\mathcal{Q}^\epsilon$ The model class $\mathcal{Q}^\epsilon$ is quite complicated. So we will begin by proving that a simpler model class $\tilde{\mathcal{Q}}^\epsilon$ is unlearnable. We will then use a set of tools developed in Appendix E to show the unlearnability of $\mathcal{Q}^\epsilon$ from the unlearnability of $\tilde{\mathcal{Q}}^\epsilon$. The simpler model class $\tilde{\mathcal{Q}}^\epsilon$ is over a simpler action spaces $\tilde{\mathcal{X}} = \{0\}, \tilde{\mathcal{Y}} = \{h, s, t\}, \tilde{\mathcal{Z}} = \{0\}$ and the same outcome space $\mathcal{B} = \{0, 1\}$. Here, the three actions h, s, t represent

Hadamard gate, Phase gate, and T gate. Each world model $\mathcal{W} = \left(\{\rho_x\}_{x\in\tilde{\mathcal{X}}}, \{\mathcal{E}_y\}_{y\in\tilde{\mathcal{Y}}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}}\right)$ in $\tilde{\mathcal{Q}}^\epsilon$ is fully specified by the following conditions,

$$\rho_0 = |0\rangle\langle 0|, \tag{F15}$$

$$\mathcal{E}_h(\rho) = \mathcal{P}_h(H\rho H^\dagger),\ \|\mathcal{P}_h - \mathcal{I}\|_\diamond \le \epsilon, \tag{F16}$$

$$\mathcal{E}_s(\rho) = \mathcal{P}_s(S\rho S^\dagger),\ \|\mathcal{P}_s - \mathcal{I}\|_\diamond \le \epsilon, \tag{F17}$$

$$\mathcal{E}_t(\rho) = \mathcal{P}_t(T\rho T^\dagger),\ \|\mathcal{P}_t - \mathcal{I}\|_\diamond \le \epsilon, \tag{F18}$$

$$\mathcal{M}_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{F19}$$

where $\mathcal{P}_h, \mathcal{P}_s, \mathcal{P}_t$ are Pauli channels. The simpler model class $\tilde{\mathcal{Q}}^\epsilon$ considers perfect zero state preparation, perfect computational basis measurement, and noisy Hadamard, phase, and T gates that are subject to Pauli noise channels. And every world model in the simpler model class $\tilde{\mathcal{Q}}^\epsilon$ is fully determines by the three Pauli noise channels $\mathcal{P}_h, \mathcal{P}_s, \mathcal{P}_t$. In the following lemma, we state the unlearnability of the simpler model class $\tilde{\mathcal{Q}}^\epsilon$.

Lemma F.2 (Gate-dependent Pauli noise is unlearnable with Hadamard+S+T gates and $|0\rangle$)**.** *Given $\epsilon > 0$. Consider a qubit system. Suppose we can prepare a perfect zero state $|0\rangle$, measure in the computational basis perfectly, and apply Hadamard gate, phase gate, and T gate, where each gate is followed by an unknown gate-dependent Pauli noise channel that is ϵ-close to the identity channel. It is impossible for any algorithm to learn the gate-dependent Pauli noise channels to arbitrarily small error. Equivalently, $\tilde{\mathcal{Q}}^\epsilon$ is unlearnable.*

The proof of Lemma F.2 is based on Proposition E.2. The proposition states that when two world models in a model class are weakly indistinguishable and are not equivalent, then the model class is unlearnable. We will identify two such world models in $\tilde{\mathcal{Q}}^\epsilon$, then apply Proposition E.2 to conclude the proof. On a high level, the existence of two weakly indistinguishable world models arises from the fact that the geometric structure for the action of Hadamard gate, phase gate, and T gate are well-aligned with Pauli noise channels, causing some noise to be indistinguishable from another.

F.4. Actions of Clifford+T on Pauli operators

We begin by illustrating the geometric structure in any quantum experiments one could perform. Every experiment under the given action spaces $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}, \tilde{\mathcal{Z}}$ is $x = 0, y_1, \ldots, y_L \in \{h, t\}, z = 0$. Now, we consider the actions of $\mathcal{E}_h$, $\mathcal{E}_s$, and $\mathcal{E}_t$ on Pauli operators,

$$\mathcal{E}_h(X) = \lambda_Z^h Z, \qquad \mathcal{E}_s(X) = \lambda_Y^s Y, \qquad \mathcal{E}_t(X) = \frac{1}{\sqrt{2}}\lambda_X^t X + \frac{1}{\sqrt{2}}\lambda_Y^t Y, \tag{F20}$$

$$\mathcal{E}_h(Y) = -\lambda_Y^h Y, \qquad \mathcal{E}_s(Y) = -\lambda_X^s X, \qquad \mathcal{E}_t(Y) = -\frac{1}{\sqrt{2}}\lambda_X^t X + \frac{1}{\sqrt{2}}\lambda_Y^t Y, \tag{F21}$$

$$\mathcal{E}_h(Z) = \lambda_X^h X, \qquad \mathcal{E}_s(Z) = \lambda_Z^s Z, \qquad \mathcal{E}_t(Z) = \lambda_Z^t Z, \tag{F22}$$

where $\lambda_X^h, \lambda_Y^h, \lambda_Z^h$ are the Pauli eigenvalues that defines the Pauli noise channel for the Hadamard gate, $\lambda_X^s, \lambda_Y^s, \lambda_Z^s$ defines the Pauli noise channel for the phase gate, and $\lambda_X^t, \lambda_Y^t, \lambda_Z^t$ defines the Pauli noise channel for the T gate. Every world model in $\mathcal{Q}^\epsilon$ is specified by the nine real values $\lambda_X^h, \lambda_Y^h, \lambda_Z^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t$.

For an experiment specified by $x = 0, y_1, \ldots, y_L \in \{h, t\}, z = 0$, the probability that the experimental outcome is 0 can be written as

$$\frac{1}{2} + \frac{1}{2}\operatorname{tr}\left(Z(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(Z)\right). \tag{F23}$$

This follows from the identities $|0\rangle\langle 0| = (I+Z)/2, |1\rangle\langle 1| = (I-Z)/2$. The probability that we obtain 1 as the experimental outcome is equal to one minus the probability for obtaining 0,

$$\frac{1}{2} - \frac{1}{2}\operatorname{tr}\left(Z(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(Z)\right). \tag{F24}$$

The actions of $\mathcal{E}_h$ and $\mathcal{E}_t$ on Pauli operators are now useful to understand the term $\operatorname{tr}(Z(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(Z))$.

F.4.a. Experiments as multiple particles traversing a graph

We represent the action of the three unitaries H, S, T on the three Pauli operators X, Y, Z as a small graph with three nodes corresponding to X, Y, Z. We can consider each experiment as particles traversing the graph. At the start of the experiment, a single particle resides on the node Z with an initial value of 1. Applying $\mathcal{E}_h$ corresponds to moving the particle from $Z \to X, Y \to Y, X \to Z$, and the value of the particle will be multiplied by $\lambda_X^h, -\lambda_Y^h, \lambda_Z^h$ accordingly. Applying $\mathcal{E}_s$ corresponds to moving the particle from $Y \to X, X \to Y, Z \to Z$, and the value of the particle will be multiplied by $-\lambda_X^s, \lambda_Y^s, \lambda_Z^s$ accordingly. Applying $\mathcal{E}_t$ corresponds to a more complicated action. If the particle resides on X, then the particle will split into two particles: one on X and one on Y. The value of the duplicated particle on X and Y will be equal to the value of the original particle multiplied by $\frac{1}{\sqrt{2}}\lambda_X^t$ and $\frac{1}{\sqrt{2}}\lambda_Y^t$ accordingly. Similarly, if the particle resides on Y, the particle will split into two particles on X and Y. The value of the duplicated particle on X and Y will be equal to the value of the original particle multiplied by $-\frac{1}{\sqrt{2}}\lambda_X^t$ and $\frac{1}{\sqrt{2}}\lambda_Y^t$ accordingly. If the particle resides on Z, then the particle will stay at Z, and the value of the particle will be multiplied by λ_Z^t. After many application of the CPTP maps $\mathcal{E}_y$, there will be many particles moving on the three-node graph. The number of particles is exponential in the number of T gates applied. At the end of the experiment, we sum up the values of particles residing at Z to obtain $\operatorname{tr}(Z(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(Z))$.

F.4.b. Polynomial forms and unlearnability

By induction, we can show that the value of each particle residing at X, Y, Z can be written as

$$\lambda_X^h(\lambda_Z^h\lambda_X^h)^{k_X} \cdot f_X(\lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t), \tag{F25}$$

$$\lambda_X^h(\lambda_Z^h\lambda_X^h)^{k_Y} \cdot f_Y(\lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t), \tag{F26}$$

$$(\lambda_Z^h\lambda_X^h)^{k_Z} \cdot f_Z(\lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t), \tag{F27}$$

accordingly, where k_X, k_Y, k_Z are non-negative integers, f_X, f_Y, f_Z are monomials. We perform induction according to the dynamics of the particles from the start of the experiment to finish. In the base case, there are only one particle residing at Z with a value of 1, hence the claimed statement holds. For each induction step, the traversal/duplication rules guarantee that the above form of values is preserved. At the end of the experiment, when we sum up the values of particles residing at Z, we have the following form for the Z expectation value,

$$\operatorname{tr}(Z(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(Z)) = \sum_k (\lambda_Z^h\lambda_X^h)^k f_k(\lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t) \tag{F28}$$

$$= p(\lambda_Z^h\lambda_X^h, \lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t), \tag{F29}$$

where k sums over non-negative integers, f_k is a monomial, p is a polynomial. Together, we can see that the probability for every experimental outcome is a polynomial function in $\lambda_Z^h\lambda_X^h, \lambda_Y^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t$. Hence, when two world models have the same $\lambda_Z^h\lambda_X^h$ and the other λ's are also equal, the two world models are weakly indistinguishable.

We now show the existence of two distinct world models that are weakly indistinguishable even under the constraint that the noise is small. Recall some basic properties of Pauli channels [42]:

$$\mathcal{P}(\rho) = p_I\rho + p_X X\rho X + p_Y Y\rho Y + p_Z Z\rho Z, \tag{F30}$$

$$= \frac{1}{2}\left(\operatorname{tr}(\rho)I + \lambda_X \operatorname{tr}(X\rho)X + \lambda_Y \operatorname{tr}(Y\rho)Y + \lambda_Z \operatorname{tr}(Z\rho)Z\right), \tag{F31}$$

$$\left\|\mathcal{P}-\mathcal{P}'\right\|_\diamond = |p_I - p'_I| + |p_X - p'_X| + |p_Y - p'_Y| + |p_Z - p'_Z|, \tag{F32}$$

where (p_I, p_X, p_Y, p_Z) is a probability distribution, $\lambda_X = 1-2p_Y-2p_Z, \lambda_Y = 1-2p_X-2p_Z, \lambda_Z = 1-2p_X-2p_Y$, and $\mathcal{P}, \mathcal{P}'$ are two Pauli channels. Each world model in $\tilde{\mathcal{Q}}^\epsilon$ is specified by $\lambda_X^h, \lambda_Y^h, \lambda_Z^h, \lambda_X^s, \lambda_Y^s, \lambda_Z^s, \lambda_X^t, \lambda_Y^t, \lambda_Z^t$, or equivalently $p_X^h, p_Y^h, p_Z^h, p_X^s, p_Y^s, p_Z^s, p_X^t, p_Y^t, p_Z^t$ as $p_I = 1-p_X-p_Y-p_Z$. We consider two world models $\mathcal{W}^A, \mathcal{W}^B$ in $\tilde{\mathcal{Q}}^\epsilon$ to be defined by

$$p_X^{h,A}=0,\ p_Z^{h,A}=\epsilon,\ p_Y^{h,A}=p_X^{s,A}=p_Y^{s,A}=p_Z^{s,A}=p_X^{t,A}=p_Y^{t,A}=p_Z^{t,A}=0, \tag{F33}$$

$$p_X^{h,B}=\epsilon,\ p_Z^{h,B}=0,\ p_Y^{h,B}=p_X^{s,B}=p_Y^{s,B}=p_Z^{s,B}=p_X^{t,B}=p_Y^{t,B}=p_Z^{t,B}=0. \tag{F34}$$

The two world models correspond to having Z-error or having X-error after Hadamard gate. Using the basic properties, it is not hard to check that both $\mathcal{W}^A, \mathcal{W}^B$ belong to $\tilde{\mathcal{Q}}^\epsilon$. Furthermore, $\lambda_X^{h,A}\lambda_Z^{h,A} = \lambda_X^{h,B}\lambda_Z^{h,B}$, and all other λ's are equal. Because the two world models are distinct, using a similar proof as in Appendix F.2, we can show that $\mathcal{W}^A \not\equiv \mathcal{W}^B$, i.e., no unitary or anti-unitary transformation exists that relates $\mathcal{W}^A$ and $\mathcal{W}^B$. Hence, $\mathcal{W}^A, \mathcal{W}^B$ are two world models that are weakly indistinguishable but are not equivalent. Using Proposition E.2, we conclude that $\tilde{\mathcal{Q}}^\epsilon$ is unlearnable.

F.5. Unlearnability of simple model class implies unlearnability of target model class

We are now ready to prove Theorem F.1, i.e., the model class $\mathcal{Q}^\epsilon$ is unlearnable. The basic structure of the proof is the following. We first utilize the equivalence of learnability after adding some actions, stated in Proposition E.3, E.4, and E.6, to show that a model class $\mathcal{R}^\epsilon$, which is a subset of $\mathcal{Q}^\epsilon$, is unlearnable. Then, we can the monotonicity of unlearnability, stated in Proposition E.1, to show that $\mathcal{Q}^\epsilon$ is unlearnable because $\mathcal{R}^\epsilon \subseteq \mathcal{Q}^\epsilon$ is unlearnable.

F.5.a. Composing all Clifford gates

We begin with the equivalence of learnability after adding composite CPTP maps stated in Proposition E.4. We will compose the two actions $h, s \in \tilde{\mathcal{Y}} = \{h, s, t\}$, which corresponds to Hadamard and phase gate subject to gate-dependent Pauli noise; see $\mathcal{E}_h$ and $\mathcal{E}_s$ in Equation (F16) and (F17). Because Hadamard and phase gates form a universal gate set for the Clifford group, we can construct any Clifford unitary C from a composition of Hadamard gate H and phase gate S. Let L^* be the required sequence length to generate every element in the Clifford unitary group. For any Clifford unitary C, we know that $f_C(P) = CPC^\dagger$ is equal to a Pauli operator (up to a phase of ± 1) for any Pauli operator $P \in \{X, Y, Z\}$. Furthermore, up to the phase, f_C is a permutation function over $\{X, Y, Z\}$. Hence, for any Clifford unitary C, Pauli channel $\mathcal{P}$, and quantum state ρ,

$$C\mathcal{P}(\rho)C^\dagger = \mathcal{Q}(C\rho C^\dagger), \tag{F35}$$

where $\mathcal{Q}$ is a different Pauli channel satisfying $\|\mathcal{Q}-\mathcal{I}\|_\diamond = \|\mathcal{P}-\mathcal{I}\|_\diamond$.

Now, we compose new CPTP maps based on the model class $\tilde{\mathcal{Q}}^{\epsilon^*}$, where ϵ^* will be chosen to be small enough later. The consideration of ϵ^* is needed to ensure that all physical operations have a small enough error. Consider a world model $\mathcal{W} = \left(\{\rho_x\}_{x\in\tilde{\mathcal{X}}}, \{\mathcal{E}_y\}_{y\in\tilde{\mathcal{Y}}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}}\right)$ in $\tilde{\mathcal{Q}}^{\epsilon^*}$. For every Clifford unitary C, there exists $y_1, \ldots, y_L \in \{h, s\}$ with $L \le L^*$, such that the composition of the Hadamard and phase gate generates the Clifford unitary C. Using the commutation relation between Pauli channel and Clifford unitary in Equation (F35) shows that

$$(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(\rho) = (\mathcal{Q}_L \circ \ldots \circ \mathcal{Q}_1)(C\rho C^\dagger), \tag{F36}$$

where $\mathcal{Q}_1, \ldots, \mathcal{Q}_L$ are Pauli channels, and $\|\mathcal{Q}_\ell - \mathcal{I}\| \le \epsilon^*, \forall \ell \in \{1, \ldots, L\}$. It is not hard to check that the composition of Pauli channel is still a Pauli channel, hence $(\mathcal{Q}_L \circ \ldots \circ \mathcal{Q}_1)$ is a Pauli channel. Using telescoping sum and triangle inequality, we have

$$\|\mathcal{Q}_L \circ \ldots \circ \mathcal{Q}_1 - \mathcal{I}\|_\diamond \le \|(\mathcal{Q}_L - \mathcal{I}) \circ \mathcal{Q}_{L-1} \circ \ldots \circ \mathcal{Q}_1\|_\diamond + \|(\mathcal{Q}_{L-1} \circ \ldots \circ \mathcal{Q}_1) - \mathcal{I}\|_\diamond \tag{F37}$$

$$\leq \|\mathcal{Q}_L - \mathcal{I}\|_\diamond + \|(\mathcal{Q}_{L-1} \circ \ldots \circ \mathcal{Q}_1) - \mathcal{I}\|_\diamond \tag{F38}$$
$$\leq \epsilon^* + \|(\mathcal{Q}_{L-1} \circ \ldots \circ \mathcal{Q}_1) - \mathcal{I}\|_\diamond \tag{F39}$$
$$\leq 2\epsilon^* + \|(\mathcal{Q}_{L-2} \circ \ldots \circ \mathcal{Q}_1) - \mathcal{I}\|_\diamond \leq \ldots \leq L\epsilon^* \leq L^*\epsilon^*. \tag{F40}$$

We add in a new action for each element in the Clifford group excluding Hadamard and phase gate H, S. After adding these new actions, our action space for CPTP maps has expanded from $\tilde{\mathcal{Y}} = \{h, s, t\}$ to $\mathcal{Y} = \mathcal{C} \cup \{T\}$, a union of the Clifford group and the T gate $\{T\}$. We call the new model class with these additional CPTP maps $\tilde{\mathcal{R}}^{\epsilon^*}$. From Proposition E.4, we know that adding these new actions do not affect the learnability. Hence $\tilde{\mathcal{R}}^{\epsilon^*}$ is unlearnable. To recap, $\tilde{\mathcal{R}}^{\epsilon^*}$ is a model class over $\tilde{\mathcal{X}} = \{0\}, \mathcal{Y} = \{y_U\}_{U \in \mathcal{C} \cup \{T\}}, \mathcal{Z} = \{0\}, \mathcal{B} = \{0, 1\}$, where $\mathcal{C}$ is the Clifford group. Every world model $\mathcal{W} = \left(\{\rho_x\}_{x \in \tilde{\mathcal{X}}}, \{\mathcal{E}_y\}_{y \in \mathcal{Y}}, \{\mathcal{M}_z\}_{z \in \mathcal{Z}}\right)$ in $\tilde{\mathcal{R}}^{\epsilon^*}$ satisfies the following condition,

$$\rho_0 = |0\rangle\langle 0|, \tag{F41}$$
$$\mathcal{E}_{y_U}(\rho) = \mathcal{P}_U(U\rho U^\dagger),\ \|\mathcal{P}_U - \mathcal{I}\|_\diamond \leq L^*\epsilon^*, \qquad \forall U \in \mathcal{C} \cup \{T\}, \tag{F42}$$
$$\mathcal{M}_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{F43}$$

where $\mathcal{P}_U, \forall U \in \mathcal{C} \cup \{T\}$ are Pauli channels. However, $\tilde{\mathcal{R}}^{\epsilon^*}$ does not contain all the world models that satisfy the above conditions.

Remark 2. The same argument works if we add in any unitary rotation about the Z-axis, i.e., any matrix diagonal in the computational basis. Under the Z-axis rotation, each particle on the X and Y node split into two particles on X and Y, and particles on the Z node stay at the Z node. The polynomial forms would still be degenerate when we add an arbitrary number of unitary rotation about the Z-axis. Therefore the gate-dependent Pauli noise is still unlearnable with Hadamard gate, phase gate, and an arbitrary number of unitary rotation about the Z-axis.

F.5.b. Composing all quantum states

The next step in the proof is to use the equivalence of learnability after adding composite initial states and adding mixture of initial states given in Proposition E.3 and E.6. We compose the zero initial state $|0\rangle\langle 0|$ with the Clifford+T gates to generate all pure states up to some small errors. Then, we can add mixture of the pure states to generate all quantum states up to some small errors. We consider L^s to be the minimum integer such that for all pure states $|\psi\rangle$, there exists a sequence $U_1, \ldots, U_\ell$ consisting of Clifford gates and T gate with $\ell \leq L^s$ and

$$\left\| |\psi\rangle\langle\psi| - (U_\ell \ldots U_1)|0\rangle\langle 0|(U_1 \ldots U_\ell)^\dagger \right\|_1 \leq \epsilon/2. \tag{F44}$$

From Solovay-Kitaev theorem [13] and follow-up works [25, 43], $L^s = \mathcal{O}(\log(1/\epsilon))$. From Proposition E.3, the model class with the additional states is constructed by specifying the set of additional actions Ξ, a constant L, and a function f with $f(\xi) = (x, y_1, \ldots, y_\ell)$. Here, we consider $\Xi = \{|\psi\rangle\}_{|\psi\rangle:\text{pure state}}$ to be the space of all pure states, $L = L^s$, and we define f to map a pure state $|\psi\rangle$ to $(0, y_{U_1}, \ldots, y_{U_\ell})$, where $U_1, \ldots, U_\ell$ are the unitaries for approximating the pure state $|\psi\rangle$ according to Eq. (F44). For every world model $\mathcal{W} = \left(\{\rho_x\}_{x \in \tilde{\mathcal{X}}}, \{\mathcal{E}_y\}_{y \in \mathcal{Y}}, \{\mathcal{M}_z\}_{z \in \mathcal{Z}}\right)$ in $\tilde{\mathcal{R}}^{\epsilon^*}$, we have

$$\left\| |\psi\rangle\langle\psi| - (\mathcal{E}_{y_{U_\ell}} \circ \ldots \circ \mathcal{E}_{y_{U_1}})(|0\rangle\langle 0|) \right\|_1 \leq \left\| |\psi\rangle\langle\psi| - (U_\ell \ldots U_1)|0\rangle\langle 0|(U_1 \ldots U_\ell)^\dagger \right\|_1 \tag{F45}$$
$$+ \left\| (\mathcal{E}_{y_{U_\ell}} \circ \ldots \circ \mathcal{E}_{y_{U_1}})(|0\rangle\langle 0|) - (U_\ell \ldots U_1)|0\rangle\langle 0|(U_1 \ldots U_\ell)^\dagger \right\|_1 \tag{F46}$$
$$\leq \epsilon/2 + \ell L^*\epsilon^* \leq \epsilon/2 + L^s L^*\epsilon^*, \tag{F47}$$

where the second-to-last inequality uses Eq. (F44), telescoping sum, and triangle inequality, and the last inequality uses $\ell \leq L^s$. We have now added actions associated to generating arbitrary pure states. Proposition E.3 shows that adding these pure states will maintain the unlearnability.

Now, for all mixed states, we only need to make use of the fact that any mixed state ρ can be written as $\rho = p|\psi\rangle\langle\psi| + (1-p)|\phi\rangle\langle\phi|$ for some $0 < p < 1$ and two orthogonal pure states $|\psi\rangle, |\phi\rangle$. Suppose $y_{U_1}, \ldots, y_{U_\ell}$ specifies the unitaries to approximate $|\psi\rangle$ and $y'_{U_1}, \ldots, y'_{U_{\ell'}}$ specifies the unitaries to approximate $|\phi\rangle$. Then by triangle inequality and Eq. (F47),

$$\left\| \rho - \left[p(\mathcal{E}_{y_{U_\ell}} \circ \ldots \circ \mathcal{E}_{y_{U_1}})(|0\rangle\langle 0|) + (1-p)(\mathcal{E}_{y'_{U_{\ell'}}} \circ \ldots \circ \mathcal{E}_{y'_{U_1}})(|0\rangle\langle 0|) \right] \right\|_1 \leq \epsilon/2 + L^s L^* \epsilon^*. \tag{F48}$$

Using Proposition E.6, we can add all the mixed states without altering the unlearnability. We have now create a model class $\mathcal{R}^{\epsilon^*}$ over $\mathcal{X} = \{x_\sigma\}_{\sigma:\text{state}}, \mathcal{Y} = \{y_U\}_{U\in\mathcal{C}\cup\{T\}}, \mathcal{Z} = \{0\}, \mathcal{B} = \{0,1\}$, where $\mathcal{C}$ is the Clifford group. Every world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$ in $\mathcal{R}^{\epsilon^*}$ satisfies

$$\|\rho_{x_\sigma} - \sigma\|_1 \leq \epsilon/2 + L^s L^* \epsilon^*, \qquad \forall \sigma : \text{state}, \tag{F49}$$

$$\rho_{x_{|0\rangle\langle 0|}} = |0\rangle\langle 0|, \tag{F50}$$

$$\mathcal{E}_{y_U}(\rho) = \mathcal{P}_U(U\rho U^\dagger),\ \|\mathcal{P}_U - \mathcal{I}\|_\diamond \leq L^*\epsilon^*, \qquad \forall U \in \mathcal{C}\cup\{T\}, \tag{F51}$$

$$\mathcal{M}_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}. \tag{F52}$$

Furthermore, from the fact that $\tilde{\mathcal{R}}^{\epsilon^*}$ is unlearnable, and Proposition E.3 and E.6, we have $\mathcal{R}^{\epsilon^*}$ is unlearnable. Now, if we choose $\epsilon^* = \epsilon/(2L^sL^*)$, then every world model in $\mathcal{R}^{\epsilon^*}$ is in $\mathcal{Q}^\epsilon$. Hence, using monotonicity of learnability in Proposition E.1, we have $\mathcal{Q}^\epsilon$ is unlearnable. This concludes the proof of Theorem F.1 by recalling the equivalence of Theorem F.1 and the unlearnability of $\mathcal{Q}^\epsilon$.

Remark 3. By tracing through the proof, one can see that the unavoidable noise floor for learning the gate-dependent Pauli noise channel is ϵ^*, which is of order $\epsilon/\log(1/\epsilon)$.

G. Noise and unlearnability

We give two examples of unlearnable model classes. Consider d-dimensional quantum worlds. We focus on the action spaces $\mathcal{X} = \{\sigma\}_{\sigma:\text{state}}, \mathcal{Y} = \{U\}_{U\in\text{SU}(d)}, \mathcal{Z} = \{0\}$. Consider $\epsilon > 0$ and the model class $\mathcal{S}^\epsilon$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ that consists of two world models $\mathcal{W}^A = \left(\{\rho_x^A\}_{x\in\mathcal{X}}, \{\mathcal{E}_y^A\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z^A\}_{z\in\mathcal{Z}}\right)$, where

$$\rho_\sigma^A = (1-\epsilon)\sigma + \epsilon\frac{I}{d}, \qquad \forall \sigma : \text{state}, \tag{G1}$$

$$\mathcal{E}_U^A(\rho) = U\rho U^\dagger, \qquad \forall U \in \text{SU}(d), \tag{G2}$$

$$\mathcal{M}_0^A = \{|b\rangle\langle b|\}_{b=1,\ldots,d}, \tag{G3}$$

and $\mathcal{W}^B = \left(\{\rho_x^B\}_{x\in\mathcal{X}}, \{\mathcal{E}_y^B\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z^B\}_{z\in\mathcal{Z}}\right)$, where

$$\rho_\sigma^B = \sigma, \qquad \forall \sigma : \text{state}, \tag{G4}$$

$$\mathcal{E}_U^B(\rho) = U\rho U^\dagger, \qquad \forall U \in \text{SU}(d), \tag{G5}$$

$$\mathcal{M}_0^B = \left\{(1-\epsilon)|b\rangle\langle b| + \epsilon\frac{I}{d}\right\}_{b=1,\ldots,d}, \tag{G6}$$

Verbally, the model class $\mathcal{S}^\epsilon$ considers world models where we have perfect unitaries, but the initial state or the computational basis measurement is subject to a depolarization noise of strength ϵ.

We also consider another model class $\mathcal{S}^\Omega$ that encompasses world models where there are noises in states, unitaries, and the computational basis measurement. Formally, $\mathcal{S}^\Omega$ is over the same set of action spaces $\mathcal{X} = \{\sigma\}_{\sigma:\text{state}}, \mathcal{Y} = \{U\}_{U\in\text{SU}(d)}, \mathcal{Z} = \{0\}$. And $\mathcal{S}^\Omega$ contains all world models $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$, where ρ_x is a quantum state, $\mathcal{E}_y$ is a CPTP map, and $\mathcal{M}_z$ is a POVM. It is not hard to see that $\mathcal{S}^\epsilon \subset \mathcal{S}^\Omega$. Hence, the two world models $\mathcal{W}^A, \mathcal{W}^B$ are also in $\mathcal{S}^\Omega$.

We will now begin to provide a formal proof showing that $\mathcal{S}^\epsilon$ is unlearnable. For any experiments $E = (\sigma, U_1, \ldots, U_\ell, 0)$, where $\sigma \in \mathcal{X}, U_1, \ldots, U_\ell \in \mathcal{Y}, 0 \in \mathcal{Z}$, we have

$$\text{tr}\left(M_{0b}^A(\mathcal{E}_{U_\ell}^A \circ \ldots \circ \mathcal{E}_{U_1}^A)(\rho_\sigma^A)\right) = (1-\epsilon)\langle b| U_\ell \ldots U_1 \sigma U_1^\dagger \ldots U_\ell^\dagger |b\rangle + \frac{\epsilon}{d},\ \forall b \in \{1,\ldots,d\}, \tag{G7}$$

$$\operatorname{tr}\left(M_{0b}^{B}(\mathcal{E}_{U_\ell}^{B}\circ\ldots\circ\mathcal{E}_{U_1}^{B})(\rho_\sigma^B)\right)=(1-\epsilon)\left\langle b\right|U_\ell\ldots U_1\sigma U_1^\dagger\ldots U_\ell^\dagger\left|b\right\rangle+\frac{\epsilon}{d},\ \forall b\in\{1,\ldots,d\}. \tag{G8}$$

Hence the two world models are weakly indistinguishable. It is also easy to check that the two world models are not equivalent to one another, hence they describe different physical realities. By Theorem A.8, if the two world models are equivalent, then there exists a unitary or anti-unitary transformation U such that $\rho_\sigma^B = U\rho_\sigma^A U^{-1}$ for any quantum state σ. We can use this relation to deduce that $\operatorname{tr}((\rho_\sigma^B)^2)=\operatorname{tr}((\rho_\sigma^A)^2)$, i.e., the purity of the two states $\rho_\sigma^A,\rho_\sigma^B$ must be equal, for all state σ. However, for $\epsilon>0$ and pure state $\sigma=|\psi\rangle\langle\psi|$, the purity of ρ_σ^A is less than one, but the purity of ρ_σ^B is one. Hence, the two world models $\mathcal{W}^A,\mathcal{W}^B$ are not equivalent.

Because $\mathcal{S}^\epsilon$ contains two weakly indistinguishable world models that are not equivalent to one another, $\mathcal{S}^\epsilon$ is unlearnable according to Proposition E.2. Then using Proposition E.1, the monotonicity of unlearnability, $\mathcal{S}^\Omega$ is unlearnable because $\mathcal{S}^\epsilon\subset\mathcal{S}^\Omega$.

H. Learning under gate-independent noise on Clifford gates

We give the detailed proof for Theorem 5 in this appendix.

H.1. Review on unitary design for Clifford gates

We recall the following well-known fact on the unitary design property for Clifford gates. These two properties will be used through the design and proof of the algorithm. Furthermore, we will be using a substantial amount of tensor network manipulation, where only the final simplified results are shown. Readers unfamiliar with tensor network manipulations could refer to reviews in [6, 9].

Lemma H.1 (Unitary design for Clifford gates [11, 49, 52])**.** *Consider $n>0$ to be the number qubits. Let $\mathcal{C}$ be the set of all Clifford gates over n qubits and let $d=2^n$. We have*

$$\frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}}CAC^\dagger=\frac{I}{d}\operatorname{tr}(A), \qquad \textit{(1-design)} \tag{H1}$$

$$\frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}}C^{\otimes 2}B(C^\dagger)^{\otimes 2}=\frac{1}{d^2-1}\left(I\operatorname{tr}(B)+S\operatorname{tr}(SB)-\frac{1}{d}S\operatorname{tr}(B)-\frac{1}{d}I\operatorname{tr}(SB)\right), \qquad \textit{(2-design)} \tag{H2}$$

where A is an $2^n\times 2^n$ complex matrix, B is a complex tensor living in the tensor product space of two $2^n\times 2^n$ complex matrices, S is the swap operator over the two tensor product components.

H.2. Learning noisy zero state, Clifford gate noise, and noisy basis measurement

Two sets of randomized experiments are conducted to learn about the noisy initial state $\rho_0\approx|0^n\rangle\langle 0^n|$, the Clifford gate noise $\mathcal{N}\approx\mathcal{I}$, and the noisy computational basis measurement $\mathcal{M}_0=\{M_b\}_{b\in\{0,1\}^n}$ with $M_b\approx|b\rangle\langle b|$. The first set of N_A experiments prepares ρ_0, evolves by $\mathcal{E}_C$ for a random Clifford C, and measures $\mathcal{M}_0$. The second set of N_B experiments prepares ρ_0, evolves by $\mathcal{E}_{C_1}$ for a random Clifford C_1, evolves by $\mathcal{E}_{C_2}$ for a second random Clifford C_2, and measures $\mathcal{M}_0$. We denote the two sets of experimental outcomes as

$$\left(C^{(A,i)},b^{(A,i)}\in\{0,1\}^n\right),\quad\forall i=1,\ldots,N_A, \tag{H3}$$

$$\left(C_1^{(B,i)},C_2^{(B,i)},b^{(B,i)}\in\{0,1\}^n\right),\quad\forall i=1,\ldots,N_B. \tag{H4}$$

We will also define the POVM $\mathcal{M}'=\{M'_b=\mathcal{N}^\dagger(M_b)\}_{b\in\{0,1\}^n}$ which is equivalent to applying the Clifford gate noise $\mathcal{N}$ followed by measuring the noisy computational basis measurement $\mathcal{M}_0$.

In the following, we will denote $d=2^n$, $f=\langle 0^n|\rho_0|0^n\rangle$, $g=\frac{1}{d}\sum_{b\in\{0,1\}^n}\langle b|M'_b|b\rangle$, $\mathbb{I}[A]$ to be the indicator function, i.e., $\mathbb{I}[A]=1$ if A is true and $\mathbb{I}[A]=0$ if A is false. Because f is unlearnable, we

can set f to whatever value we want. One practical choice is to set $f = 1$ since $\rho_0 \approx |0^n\rangle\langle 0^n|$. Under the assumption that $\rho_0 \approx |0^n\rangle\langle 0^n|$ and $M'_b = \mathcal{N}^\dagger(M_b) \approx M_b \approx |b\rangle\langle b|$, we can consider

$$f = \langle 0^n| \rho_0 |0^n\rangle \in [0.9, 1], \quad \langle b| M'_b |b\rangle \in [0.9, 1.1], \quad g = \frac{1}{d} \sum_{b\in\{0,1\}^n} \langle b| M'_b |b\rangle \in [0.9, 1.1]. \tag{H5}$$

This follows from the fact that the noise is bounded.

H.2.a. Learning noisy basis measurement conflated with Clifford gate noise

We now construct various estimators that characterizes the POVM $\mathcal{M}' = \{M'_b\}_{b\in\{0,1\}^n}$. We start with the simplest estimator,

$$\widehat{\mathrm{tr}(M'_b)} = \frac{1}{N_A} \sum_{i=1}^{N_A} d\, \mathbb{I}\left[b^{(A,i)} = b\right]. \tag{H6}$$

Using unitary 1-design property of Clifford gates, we have

$$\mathbb{E}\left[\widehat{\mathrm{tr}(M'_b)}\right] = \frac{1}{|\mathcal{C}|} \sum_{C\in\mathcal{C}} d\,\mathrm{tr}\left(M'_b C\rho_0 C^\dagger\right) = \mathrm{tr}(M'_b). \tag{H7}$$

Hence, we can estimate $\mathrm{tr}(M'_b)$ to arbitrarily small error with large enough N_A. The next estimator is slightly more complicated and uses $\widehat{\mathrm{tr}(M'_b)}$,

$$\widehat{M'_b} = \frac{d^2-1}{f-\frac{1}{d}} \left[\frac{1}{N_A} \sum_{i=1}^{N_A} \mathbb{I}\left[b^{(A,i)} = b\right] C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger\right] - \frac{1-\frac{f}{d}}{f-\frac{1}{d}} \widehat{\mathrm{tr}(M'_b)} I. \tag{H8}$$

Using unitary 2-design property of Clifford gates, we have

$$\mathbb{E}\left[\widehat{M'_b}\right] = \frac{1}{|\mathcal{C}|} \sum_{C\in\mathcal{C}} \frac{d^2-1}{f-\frac{1}{d}} \mathrm{tr}(M'_b C\rho_0 C^\dagger) C|0^n\rangle\langle 0^n|C^\dagger - \frac{1-\frac{f}{d}}{f-\frac{1}{d}} \mathrm{tr}(M'_b) I = M'_b. \tag{H9}$$

Hence, we can estimate the POVM element M'_b to arbitrarily small error with large enough N_A.

We will also utilize the following estimator to estimate $g = \frac{1}{d}\sum_{b\in\{0,1\}^n} \langle b| M'_b |b\rangle$.

$$\widehat{g} = \frac{1}{d} \sum_{b\in\{0,1\}^n} \langle b| \widehat{M'_b} |b\rangle = \frac{d^2-1}{df-1} \left[\frac{1}{N_A} \sum_{i=1}^{N_A} \langle b^{(A,i)}| C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger |b^{(A,i)}\rangle\right] - \frac{d-f}{df-1}. \tag{H10}$$

The second equality follows from $\sum_{b\in\{0,1\}^n} \widehat{\mathrm{tr}(M'_b)} = d$. From the equality in Eq. (H9) and linearity of expectation, we have

$$\mathbb{E}\left[\widehat{g}\right] = \frac{1}{d} \sum_{b\in\{0,1\}^n} \langle b| M'_b |b\rangle = g. \tag{H11}$$

Hence, we can estimate the scalar value g to arbitrarily small error with large enough N_A.

H.2.b. Learning noisy zero state

If g is perfectly known, then the estimator for ρ_0 is given as follows.

$$\widehat{\rho_0}^{(g*)} = \frac{d^2-1}{dg-1} \left[\frac{1}{N_A} \sum_{i=1}^{N_A} (C^{(A,i)})^\dagger |b^{(A,i)}\rangle\langle b^{(A,i)}| C^{(A,i)}\right] - \frac{d-\frac{g}{d}}{dg-1} I. \tag{H12}$$

Notice that in the noiseless setting ($g = 1$), the above estimator is exactly equal to the classical shadow representation based on randomized Clifford measurements [29]. Using the unitary 2-design property of Clifford gates, we have

$$\mathbb{E}\left[\widehat{\rho_0}^{(g*)}\right] = \frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}}\frac{d^2-1}{dg-1}\operatorname{tr}(M_b' C\rho_0 C^\dagger)C^\dagger|b\rangle\langle b|C - \frac{d-\frac{g}{d}}{dg-1}I = \rho_0. \tag{H13}$$

Hence, we can estimate ρ_0 to arbitrarily small error with large enough N_A when g is perfectly known. However, since g is estimated using $\hat{g}$, we will use the following estimator instead.

$$\widehat{\rho_0} = \frac{d^2-1}{d\hat{g}-1}\left[\frac{1}{N_A}\sum_{i=1}^{N_A}(C^{(A,i)})^\dagger|b^{(A,i)}\rangle\langle b^{(A,i)}|C^{(A,i)}\right] - \frac{d-\frac{\hat{g}}{d}}{d\hat{g}-1}I. \tag{H14}$$

We use that fact that, with large enough N_A, $\hat{g}$ can be made arbitrarily close to g. Because $g \in [0.9, 1.1]$ from Eq. (H5), with large enough N_A, $\frac{d^2-1}{d\hat{g}-1}$ and $\frac{d-\frac{\hat{g}}{d}}{d\hat{g}-1}$ can be made arbitrarily close to $\frac{d^2-1}{dg-1}$ and $\frac{d-\frac{g}{d}}{dg-1}$, respectively. Hence, with Eq. (H13), the estimator $\widehat{\rho_0}$ can be made arbitrarily close to ρ_0 with large enough N_A.

H.2.c. Learning Clifford gate noise

So far, we have only used the first set of experiments. We are now ready to learn the Clifford gate noise $\mathcal{N}$ using the second set of experiments. We will use the Choi matrix representation of a quantum channel. Recall that the Choi matrix of a channel $\mathcal{N}$ is given by

$$\Phi_{\mathcal{N}} \equiv (\mathcal{N}\otimes\mathcal{I})(|\omega\rangle\langle\omega|), \tag{H15}$$

where $|\omega\rangle = \frac{1}{\sqrt{d}}\sum_{b\in\{0,1\}^n}|b\rangle\otimes|b\rangle$ is the maximally entangled state over $2n$ qubits, $\mathcal{I}$ is the identity channel on n qubits. We first estimate the state $\mathcal{N}(I/d)$. The basic idea is to intentionally neglect $C_1^{(B,i)}$ in the data. Then ρ_0 evolved under an unknown random Clifford gate C_1 followed by the gate noise $\mathcal{N}$ is equal to the state $\mathcal{N}(I/d)$ from the unitary 1-design property of random Clifford gate. Then we can essentially use the same estimator as Eq. (H12) to learn the state $\mathcal{N}(I/d)$. Assuming g is known, then we can use the estimator,

$$\widehat{\mathcal{N}(I/d)}^{(g*)} = \frac{d^2-1}{dg-1}\left[\frac{1}{N_B}\sum_{i=1}^{N_B}(C_2^{(B,i)})^\dagger|b^{(B,i)}\rangle\langle b^{(B,i)}|C_2^{(B,i)}\right] - \frac{d-\frac{g}{d}}{dg-1}I. \tag{H16}$$

From unitary 1-design and 2-design property, we have

$$\mathbb{E}\left[\widehat{\mathcal{N}(I/d)}^{(g*)}\right] = \frac{1}{|\mathcal{C}|^2}\sum_{C_1\in\mathcal{C}}\sum_{C_2\in\mathcal{C}}\frac{d^2-1}{dg-1}\operatorname{tr}(M_b' C_2\mathcal{N}(C_1\rho_0 C_1^\dagger)C_2^\dagger)C_2^\dagger|b\rangle\langle b|C_2 - \frac{d-\frac{g}{d}}{dg-1}I \tag{H17}$$

$$= \mathcal{N}(I/d). \tag{H18}$$

Similar to learning ρ_0, we only have an estimate for g given by $\hat{g}$. Hence, the estimator we will use is

$$\widehat{\mathcal{N}(I/d)} = \frac{d^2-1}{d\hat{g}-1}\left[\frac{1}{N_B}\sum_{i=1}^{N_B}(C_2^{(B,i)})^\dagger|b^{(B,i)}\rangle\langle b^{(B,i)}|C_2^{(B,i)}\right] - \frac{d-\frac{\hat{g}}{d}}{d\hat{g}-1}I. \tag{H19}$$

With large enough N_A, $\hat{g}$ can be made arbitrarily close to g. So, from Eq (H18), the estimator $\widehat{\mathcal{N}(I/d)}$ can be made arbitrarily close to $\mathcal{N}(I/d)$ with large enough N_A and N_B.

We now present the estimator for the Choi matrix of $\mathcal{N}$. We will begin by assuming that g is perfectly known, then we will approximate g by $\widehat{g}$. The estimator $\widehat{\Phi_{\mathcal{N}}}^{(g*)}$ is defined as

$$\widehat{\Phi_{\mathcal{N}}}^{(g*)} = \frac{(d^2-1)^2}{(df-1)(dg-1)}\left[\frac{1}{N_B}\sum_{i=1}^{N_B}\left((C_2^{(B,i)})^\dagger|b^{(B,i)}\rangle\langle b^{(B,i)}|C_2^{(B,i)}\right)\otimes\left(C_1^{(B,i)}|0^n\rangle\langle 0^n|(C_1^{(B,i)})^\dagger\right)^T\right] \tag{H20}$$

$$- \frac{(d-\frac{1}{d})(d-\hat{g})}{(df-1)(dg-1)} \ [I\otimes I] - \frac{1-\frac{f}{d}}{f-\frac{1}{d}} \ \left[\widehat{\mathcal{N}(I/d)}^{(g*)}\otimes I\right]. \tag{H21}$$

Using unitary 2-design property of Clifford gates and Eq. (H18), we have

$$\mathbb{E}\left[\widehat{\Phi_{\mathcal{N}}}^{(g*)}\right] = \frac{1}{|\mathcal{C}|^2}\sum_{C_1\in\mathcal{C}}\sum_{C_2\in\mathcal{C}} \frac{(d^2-1)^2}{(df-1)(g-1)} \operatorname{tr}(M_b' C_2\mathcal{N}(C_1\rho_0 C_1^\dagger)C_2^\dagger) \tag{H22}$$

$$\times \left(C_2^\dagger|b\rangle\langle b|C_2\right)\otimes\left(C_1|0^n\rangle\langle 0^n|C_1^\dagger\right)^T \tag{H23}$$

$$- \frac{(d-\frac{1}{d})(1-\frac{g}{d^2})}{(df-1)(g-1)} I\otimes I - \frac{1-\frac{f}{d}}{f-\frac{1}{d}}\mathcal{N}(I/d)\otimes I \tag{H24}$$

$$= \Phi_{\mathcal{N}}. \tag{H25}$$

Similar to before, we only have an estimate for g given by $\hat{g}$, so we will instead use the following estimator, i.e., replacing all g by $\hat{g}$ and $\widehat{\mathcal{N}(I/d)}^{(g*)}$ by $\widehat{\mathcal{N}(I/d)}$.

$$\widehat{\Phi_{\mathcal{N}}} = \frac{(d^2-1)^2}{(df-1)(d\hat{g}-1)}\left[\frac{1}{N_B}\sum_{i=1}^{N_B}\left((C_2^{(B,i)})^\dagger|b^{(B,i)}\rangle\langle b^{(B,i)}|C_2^{(B,i)}\right)\otimes\left(C_1^{(B,i)}|0^n\rangle\langle 0^n|(C_1^{(B,i)})^\dagger\right)^T\right] \tag{H26}$$

$$- \frac{(d-\frac{1}{d})(d-\hat{g})}{(df-1)(d\hat{g}-1)} \ [I\otimes I] - \frac{1-\frac{f}{d}}{f-\frac{1}{d}} \ \left[\widehat{\mathcal{N}(I/d)}\otimes I\right]. \tag{H27}$$

With large enough N_A, N_B, $\hat{g}$ and $\widehat{\mathcal{N}(I/d)}$ can be made arbitrarily close to g and $\mathcal{N}(I/d)$. So, from Eq (H25), the estimator $\widehat{\Phi_{\mathcal{N}}}$ can be made arbitrarily close to $\Phi_{\mathcal{N}}$ with large enough N_A and N_B. With the estimator for Choi matrix $\widehat{\Phi_{\mathcal{N}}}$, we can obtain the estimator $\widehat{\mathcal{N}}$ for the original CPTP map $\mathcal{N}$ by applying the linear invertible transformation between Choi matrix and CPTP map.

H.2.d. Learning noisy basis measurement

After learning $\Phi_{\mathcal{N}}$ through the estimator $\widehat{\Phi_{\mathcal{N}}}$, we can obtain the noisy computational basis measurement $\mathcal{M}_0 = \{M_b\}_{b\in\{0,1\}^n}$ by considering the following estimator.

$$\widehat{M_b} = \left(\widehat{\mathcal{N}^\dagger}\right)^{-1}\widehat{M_b'}, \quad \forall b\in\{0,1\}^n. \tag{H28}$$

Note that $\widehat{M_b'}$ can be made arbitrarily close to $M_b' = \mathcal{N}^\dagger(M_b)$ and $\widehat{\mathcal{N}}$ can be made arbitrarily close to $\mathcal{N}$. Because $\mathcal{N}$ is assumed to be close to the identity, the difference between $\left(\widehat{\mathcal{N}^\dagger}\right)^{-1}$ and $\left(\mathcal{N}^\dagger\right)^{-1}$ can be made arbitrarily small by increasing N_A, N_B. Hence, $\widehat{M_b}$ can be arbitrarily close to M_b with large enough N_A, N_B.

H.3. Learning any state, CPTP map, and POVM

After learning $\rho_0, \mathcal{N}, \mathcal{M}_0$, we present the learning of any physical operation in the system. The procedures are very similar to the previous subsection.

H.3.a. Learning POVM

Given any POVM $\mathcal{M}_z = \{M_{zb}\}_{b\in\mathcal{B}}$, where $\mathcal{B}$ is a set denoting all the possible outcomes. We can learn M_{zb} by conducting the following randomized experiments for N_z times: prepare ρ_0, evolve under $\mathcal{E}_C$ for a random Clifford gate C, measure $\mathcal{M}_z$. We denote the sets of experimental outcomes as

$$\left(C^{(z,i)}, b^{(z,i)}\in\mathcal{B}\right), \quad \forall i = 1,\ldots,N_z. \tag{H29}$$

The procedure is very similar to that for learning $\mathcal{M}_0$. First, we consider the POVM $\mathcal{M}'_z = \{M'_{zb} = \mathcal{N}^\dagger(M_z b)\}_{b\in\mathcal{B}}$, which conflates $\mathcal{M}_z$ with the Clifford gate noise $\mathcal{N}$. We can learn $\mathcal{M}'_z$ using the following estimators based on the data we obtained.

$$\widehat{\mathrm{tr}(M'_{zb})} = \frac{1}{N_z}\sum_{i=1}^{N_z} d\,\mathbb{I}\left[b^{(z,i)} = b\right], \tag{H30}$$

$$\widehat{M'_{zb}} = \frac{d^2-1}{f-\frac{1}{d}}\left[\frac{1}{N_z}\sum_{i=1}^{N_z}\mathbb{I}\left[b^{(z,i)}=b\right]C^{(z,i)}|0^n\rangle\langle 0^n|(C^{(z,i)})^\dagger\right] - \frac{1-\frac{f}{d}}{f-\frac{1}{d}}\widehat{\mathrm{tr}(M'_{zb})}I. \tag{H31}$$

Using the same analysis as Appendix H.2.a, we can show that $\widehat{M'_{zb}}$ can be made arbitrarily close to M'_{zb} with large enough N_z. Then similar to Appendix H.2.d, we can learn M_{zb} using the following estimator after obtaining $\hat{\mathcal{N}}$ from the algorithms presented in Appendix H.2.c.

$$\widehat{M_{zb}} = \left(\hat{\mathcal{N}}^\dagger\right)^{-1}\widehat{M'_{zb}}, \quad \forall b\in\mathcal{B}. \tag{H32}$$

With large enough N_A, N_B, N_z, we can make $\widehat{M_{zb}}$ arbitrarily close to M_{zb}.

H.3.b. Learning states

Given any state ρ_x. We can learn ρ_x by conducting the following randomized experiments for N_x times: prepare ρ_x, evolve under $\mathcal{E}_C$ for a random Clifford gate C, measure $\mathcal{M}_0$. We denote the sets of experimental outcomes as

$$\left(C^{(x,i)}, b^{(x,i)}\in\{0,1\}^n\right), \quad \forall i = 1,\ldots,N_x. \tag{H33}$$

We can obtain an estimate for ρ_x using the following formula.

$$\widehat{\rho_x} = \frac{d^2-1}{d\hat{g}-1}\left[\frac{1}{N_x}\sum_{i=1}^{N_x}(C^{(x,i)})^\dagger|b^{(x,i)}\rangle\langle b^{(x,i)}|C^{(x,i)}\right] - \frac{d-\frac{\hat{g}}{d}}{d\hat{g}-1}I. \tag{H34}$$

With large enough N_A, N_B, N_x, we can make $\widehat{\rho_x}$ arbitrarily close to ρ_x. The analysis is the same as that in Appendix H.2.b.

H.3.c. Learning CPTP maps

Given any CPTP map $\mathcal{E}_y$. We can learn $\mathcal{E}_y$ by conducting the following randomized experiments for N_y times: prepare ρ_0, evolve under $\mathcal{E}_{C_1}$ for a random Clifford gate C_1, evolve under $\mathcal{E}_y$, evolve under $\mathcal{E}_{C_2}$ for a different random Clifford gate C_2, measure $\mathcal{M}_0$. We denote the sets of N_y experimental outcomes as follows.

$$\left(C_1^{(y,i)}, C_2^{(y,i)}, b^{(y,i)}\in\{0,1\}^n\right), \quad \forall i = 1,\ldots,N_y. \tag{H35}$$

We first learn the quantum process $\mathcal{E}'_y = \mathcal{E}_y\circ\mathcal{N}$, which conflates $\mathcal{E}_y$ with the Clifford gate noise $\mathcal{N}$. Following the same analysis as Appendix H.2.c but replacing $\mathcal{N}$ by $\mathcal{E}'_y$, we define the following estimators.

$$\widehat{\mathcal{E}'_y(I/d)} = \frac{d^2-1}{d\hat{g}-1}\left[\frac{1}{N_y}\sum_{i=1}^{N_y}(C_2^{(u,i)})^\dagger|b^{(y,i)}\rangle\langle b^{(y,i)}|C_2^{(y,i)}\right] - \frac{d-\frac{\hat{g}}{d}}{d\hat{g}-1}I. \tag{H36}$$

$$\widehat{\Phi_{\mathcal{E}'_y}} = \frac{(d^2-1)^2}{(df-1)(d\hat{g}-1)}\left[\frac{1}{N_y}\sum_{i=1}^{N_y}\left((C_2^{(y,i)})^\dagger|b^{(y,i)}\rangle\langle b^{(y,i)}|C_2^{(y,i)}\right)\otimes\left(C_1^{(y,i)}|0^n\rangle\langle 0^n|(C_1^{(y,i)})^\dagger\right)^T\right] \tag{H37}$$

$$- \frac{(d-\frac{1}{d})(d-\hat{g})}{(df-1)(d\hat{g}-1)} \left[I\otimes I\right] - \frac{1-\frac{f}{d}}{f-\frac{1}{d}} \left[\widehat{\mathcal{N}(I/d)}\otimes I\right]. \tag{H38}$$

With large enough N_A, N_B, N_y, we can make $\widehat{\Phi_{\mathcal{E}'_y}}$ arbitrarily close to $\Phi_{\mathcal{E}'_y}$, the Choi matrix for the CPTP map $\mathcal{E}'_y$. Hence, we can obtain the estimator $\widehat{\mathcal{E}}'_y$ for $\mathcal{E}'_y$ by transforming the Choi matrix back to the CPTP map. Now we simply need to invert the conflation with $\mathcal{N}$ by considering

$$\widehat{\mathcal{E}}_y = \widehat{\mathcal{E}}'_y \circ (\widehat{\mathcal{N}})^{-1}. \tag{H39}$$

Because the Clifford gate noise $\mathcal{N}$ is assumed to be close to the identity, the difference between $\left(\widehat{\mathcal{N}}^\dagger\right)^{-1}$ and $\left(\mathcal{N}^\dagger\right)^{-1}$ can be made arbitrarily small by increasing N_A, N_B. Therefore, we can make the difference between $\widehat{\mathcal{E}}_y$ and $\mathcal{E}_y$ arbitrarily small with large enough N_A, N_B, N_y.

H.4. Sample complexity for learning Clifford gate noise

All the previous analyses could be equipped with rigorous convergence guarantee using concentration inequalities similar to quantum state/process tomography based on randomized measurements [22, 45]. As an example, we present the sample complexity for learning Clifford gate noise $\mathcal{N}$ with $\mathcal{N}(I) = I$. The reconstruction of $\mathcal{N}$ under the condition $\mathcal{N}(I) = I$ was previously studied in [26, 33, 41], where learning algorithms based on interleaved randomized benchmarking [36] have been devised. We will show that the proposed algorithm is much more efficient than the best known sample complexity of $\mathcal{O}(d^8)$ in [26] for learning the Choi matrix $\Phi_{\mathcal{N}}$ up to a constant error in Hilbert-Schmidt norm.

Under the assumption that $\mathcal{N}(I) = I$, we do not need to estimate $\mathcal{N}(I/d)$ because $\mathcal{N}(I/d) = I/d$. Hence, the estimator in Eq. (H27) simplifies to

$$\widehat{\Phi_{\mathcal{N}}} = \frac{(d^2-1)^2}{(df-1)(d\hat{g}-1)} \left[\frac{1}{N_B}\sum_{i=1}^{N_B} \left((C_2^{(B,i)})^\dagger |b^{(B,i)}\rangle\langle b^{(B,i)}| C_2^{(B,i)}\right) \otimes \left(C_1^{(B,i)}|0^n\rangle\langle 0^n|(C_1^{(B,i)})^\dagger\right)^T\right] \tag{H40}$$

$$- \frac{(d-\frac{1}{d})(d-\hat{g})}{(df-1)(d\hat{g}-1)} \left[I\otimes I\right] - \frac{1-\frac{f}{d}}{df-1} \left[I\otimes I\right]. \tag{H41}$$

Furthermore, using the definition of $\hat{g}$ in Eq. (H10), we see that

$$(df-1)(d\hat{g}-1) = (d^2-1)\left(d\left[\frac{1}{N_A}\sum_{i=1}^{N_A}\langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger\,|b^{(A,i)}\rangle\right] - 1\right) \tag{H42}$$

is independent of f. Furthermore, we can simplify the coefficients for $I\otimes I$ using

$$- \frac{(d-\frac{1}{d})(d-\hat{g})}{(df-1)(d\hat{g}-1)} - \frac{1-\frac{f}{d}}{df-1} = -\frac{(d-\frac{1}{d})^2}{(df-1)(d\hat{g}-1)} + \frac{1}{d^2}, \tag{H43}$$

which is also independent of f because $(df-1)(d\hat{g}-1)$ is independent of f. We can now rewrite the estimator $\widehat{\Phi_{\mathcal{N}}}$ as the following f-independent expression,

$$\widehat{\Phi_{\mathcal{N}}} = (d^2-1)\frac{\frac{1}{N_B}\sum_{i=1}^{N_B}\left((C_2^{(B,i)})^\dagger |b^{(B,i)}\rangle\langle b^{(B,i)}| C_2^{(B,i)}\right)\otimes\left(C_1^{(B,i)}|0^n\rangle\langle 0^n|(C_1^{(B,i)})^\dagger\right)^T}{d\left[\frac{1}{N_A}\sum_{i=1}^{N_A}\langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger\,|b^{(A,i)}\rangle\right]-1} \tag{H44}$$

$$- (d^2-1)\frac{(I/d)\otimes(I/d)}{d\left[\frac{1}{N_A}\sum_{i=1}^{N_A}\langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger\,|b^{(A,i)}\rangle\right]-1} + (I/d)\otimes(I/d). \tag{H45}$$

We will now present the algorithm for estimating $\Phi_{\mathcal{N}}$ under the Pauli basis. Consider two n-qubit Pauli operators $P, Q \in \{I, X, Y, Z\}^{\otimes n}$. We would estimate $\mathrm{tr}((P\otimes Q)\Phi_{\mathcal{N}})$ based on the above expression for $\widehat{\Phi_{\mathcal{N}}}$. The estimation procedure is separated into three cases.

H.4.a. $P = I$ and $Q = I$

This is the simplest case. Because $\Phi_{\mathcal{N}}$ is a quantum state, we have $\mathrm{tr}(\Phi_{\mathcal{N}}) = 1$. Hence, we can obtain a perfect estimate for $\mathrm{tr}((P \otimes Q)\Phi_{\mathcal{N}})$ as it is always one.

H.4.b. Exactly one of P and Q is equal to I

This is also a simple case. Recall that a non-identity Pauli operator has trace equal to zero. Then, using the definition of Choi matrix $\Phi_{\mathcal{N}}$ and $\mathcal{N}(I) = I$, we have

$$\mathrm{tr}((P \otimes Q)\Phi_{\mathcal{N}}) = 0. \tag{H46}$$

Hence, we can obtain a perfect estimate for $\mathrm{tr}((P \otimes Q)\Phi_{\mathcal{N}})$ as it is always zero.

H.4.c. $P \neq I$ and $Q \neq I$

Using the fact that a non-identity Pauli operator has trace equal to zero, we have

$$\mathrm{tr}((P \otimes Q)\widehat{\Phi_{\mathcal{N}}}) = (d^2-1)\frac{\frac{1}{N_B}\sum_{i=1}^{N_B} \langle b^{(B,i)}|\, C_2^{(B,i)} P (C_2^{(B,i)})^\dagger \,|b^{(B,i)}\rangle \langle 0^n|\, (C_1^{(B,i)})^\dagger Q^T C_1^{(B,i)} \,|0^n\rangle}{d\left[\frac{1}{N_A}\sum_{i=1}^{N_A} \langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger \,|b^{(A,i)}\rangle\right] - 1}. \tag{H47}$$

We will directly use the above formula to estimate $\mathrm{tr}((P \otimes Q)\Phi_{\mathcal{N}})$.

To analyze the error in the above estimator, we separately consider the convergence of the numerator and the denominator. We begin with the denominator,

$$\mathcal{Y} = d\left[\frac{1}{N_A}\sum_{i=1}^{N_A} \langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger \,|b^{(A,i)}\rangle\right] - 1 \tag{H48}$$

$$= \frac{1}{N_A}\sum_{i=1}^{N_A}\left[d\,\langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger \,|b^{(A,i)}\rangle - 1\right]. \tag{H49}$$

Lemma H.2 (Concentration for the denominator)**.** *Fix $\epsilon > 0$. Given $N_A = \mathcal{O}(1/\epsilon^2)$. With probability at least 0.99, we have*

$$\left|\mathcal{Y} - \frac{(df-1)(dg-1)}{(d-1)(d+1)}\right| < \epsilon, \tag{H50}$$

where $\frac{(df-1)(dg-1)}{(d-1)(d+1)} \in [0.21, 1.21]$ from Eq. (H5).

Proof. Let $\mathcal{Y}_i = d\,\langle b^{(A,i)}|\, C^{(A,i)}|0^n\rangle\langle 0^n|(C^{(A,i)})^\dagger \,|b^{(A,i)}\rangle - 1$. From unitary 2-design of random Clifford gate, we have the following identity,

$$\mathbb{E}[\mathcal{Y}_i] = \frac{(df-1)(dg-1)}{(d-1)(d+1)}. \tag{H51}$$

Then, using the unitary 3-design property of random Clifford gate [49, 52] and the conditions in Eq. (H5), we have

$$\mathrm{Var}[\mathcal{Y}_i] \leq \frac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}}\sum_{b\in\{0,1\}^n} \mathrm{tr}(M_b' C\rho_0 C^\dagger) d^2 (\langle b|\, C|0^n\rangle\langle 0^n|C^\dagger \,|b\rangle)^2 = \mathcal{O}(1). \tag{H52}$$

The claim then follows from $\mathcal{Y} = \frac{1}{N_A}\sum_{i=1}^{N_A}\mathcal{Y}_i$ and Chebyshev's inequality. □

Next, we can analyze the numerator,

$$\mathcal{X} = \frac{1}{N_B}\sum_{i=1}^{N_B}(d^2-1)\left\langle b^{(B,i)}\right| C_2^{(B,i)} P (C_2^{(B,i)})^\dagger \left|b^{(B,i)}\right\rangle \left\langle 0^n\right| (C_1^{(B,i)})^\dagger Q^T C_1^{(B,i)} \left|0^n\right\rangle . \tag{H53}$$

Lemma H.3 (Concentration for the numerator)**.** *Fix* $0.5 > \epsilon, \delta > 0$. *Given* $N_B = \mathcal{O}(d^2\log(1/\delta)/\epsilon^2)$. *With probability at least* $1-\delta$,

$$\left|\mathcal{X} - \frac{(df-1)(dg-1)}{(d-1)(d+1)}\operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})\right| < \epsilon, \tag{H54}$$

where $\frac{(df-1)(dg-1)}{(d-1)(d+1)}\operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}}) \in [0.21, 1.21]$ *from Eq.* (H5).

Proof. Let $\mathcal{X}_i = (d^2-1)\left\langle b^{(B,i)}\right| C_2^{(B,i)} P (C_2^{(B,i)})^\dagger \left|b^{(B,i)}\right\rangle \left\langle 0^n\right| (C_1^{(B,i)})^\dagger Q^T C_1^{(B,i)} \left|0^n\right\rangle$. From unitary 2-design of random Clifford gate and the fact that non-identity Pauli has trace zero, we have,

$$\mathbb{E}[\mathcal{X}_i] = \frac{(df-1)(dg-1)}{(d-1)(d+1)}\operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}}). \tag{H55}$$

Because $C_2^{(B,i)}$ is a random Clifford gate, $C_2^{(B,i)} P (C_2^{(B,i)})^\dagger$ is proportional to a random non-identity Pauli $\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}$. If $C_2^{(B,i)} P (C_2^{(B,i)})^\dagger$ is not proportional to a Pauli-Z operator $\{I,Z\}^{\otimes n}$, then $\mathcal{X}_i = 0$. Similarly $(C_1^{(B,i)})^\dagger Q^T C_1^{(B,i)}$ is proportional to a random non-identity Pauli. And, $C_1^{(B,i)} Q^T (C_1^{(B,i)})^\dagger$ is not proportional to a Pauli-Z operator $\{I,Z\}^{\otimes n}$, then $\mathcal{X}_i = 0$. Because C_1, C_2 are independent random gates, we have $\mathcal{X}_i \neq 0$ with probability at most

$$\frac{(d-1)}{(d^2-1)} \times \frac{(d-1)}{(d^2-1)} = \frac{1}{(d+1)^2}. \tag{H56}$$

Furthermore, we have $|\mathcal{X}_i| \leq (d^2-1)$ with probability one. Therefore, we have

$$\operatorname{Var}[\mathcal{X}_i] \leq \mathbb{E}[\mathcal{X}_i^2] \leq \frac{1}{(d+1)^2} \times (d^2-1)^2 = (d-1)^2 \leq d^2. \tag{H57}$$

From Bernstein's inequality and the definition that $\mathcal{X} = \frac{1}{N_B}\sum_{i=1}^{N_B}\mathcal{X}_i$, we conclude the proof. □

We can now establish the following statement.

Lemma H.4 (Combine numerator and denominator)**.** *Given* $0 < \epsilon < 0.1$. *Assume Eq.* (H50) *and* (H54) *both hold. Then*

$$\left|\frac{\mathcal{X}}{\mathcal{Y}} - \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})\right| < 155\epsilon. \tag{H58}$$

Proof. The proof follows from the following analysis.

$$\left|\frac{\mathcal{X}}{\mathcal{Y}} - \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})\right| \leq \left|\frac{\mathcal{X}}{\mathcal{Y}} - \frac{\mathcal{X}}{\frac{(df-1)(dg-1)}{(d-1)(d+1)}}\right| + \left|\frac{\mathcal{X}}{\frac{(df-1)(dg-1)}{(d-1)(d+1)}} - \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})\right| \tag{H59}$$

$$\leq (1.21+\epsilon)\left|\frac{1}{\mathcal{Y}} - \frac{1}{\frac{(df-1)(dg-1)}{(d-1)(d+1)}}\right| \tag{H60}$$

$$+ 5\left|\mathcal{X} - \frac{(df-1)(dg-1)}{(d-1)(d+1)}\operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})\right| \tag{H61}$$

$$< 1.5\times 100\epsilon + 5\epsilon = 155\epsilon. \tag{H62}$$

The first line uses triangle inequality. The second inequality uses the fact that

$$\frac{(df-1)(dg-1)}{(d-1)(d+1)}, \frac{(df-1)(dg-1)}{(d-1)(d+1)} \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}}) \in [0.21, 1.21]. \tag{H63}$$

The third inequality uses Eq. (H63), the condition that $0 < \epsilon < 0.1$, and

$$\left|\frac{1}{a} - \frac{1}{b}\right| < \frac{1}{\min(a,b)^2} \times |a-b|, \forall a, b > 0. \tag{H64}$$

This concludes the proof. □

H.4.d. Final result

Recall the following Pauli basis representation of an $2n$-qubit quantum state,

$$\Phi_{\mathcal{N}} = \sum_{P,Q\in\{I,X,Y,Z\}^{\otimes n}} \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}}) \frac{(P\otimes Q)}{d^2}. \tag{H65}$$

We can learn the coefficients $\operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})$ using the above strategy. We denote the estimated coefficients as $\hat{o}(\Phi_{\mathcal{N}}, P, Q) \in \mathbb{R}$. We can then obtain a reconstruction for $\Phi_{\mathcal{N}}$ as

$$\tilde{\Phi}_{\mathcal{N}}(N_A, N_B) = \sum_{P,Q\in\{I,X,Y,Z\}^{\otimes n}} \hat{o}(\Phi_{\mathcal{N}}, P, Q) \frac{(P\otimes Q)}{d^2}. \tag{H66}$$

We can now combine the previous results to show the sample complexity to learn the Choi matrix of the Clifford gate noise $\Phi_{\mathcal{N}}$ up to ϵ error in Hilbert Schmidt norm.

Theorem H.5 (Sample complexity for learning Clifford gate noise)**.** *Given $0 < \epsilon < 0.1d$. Assume the noise is bounded as stated in Eq. (H5). With $N_A = \mathcal{O}(d^2/\epsilon^2)$ and $N_B = \mathcal{O}(d^4 \log(d)/\epsilon^2)$,*

$$\left\|\tilde{\Phi}_{\mathcal{N}}(N_A, N_B) - \Phi_{\mathcal{N}}\right\|_{\mathrm{HS}} < \epsilon, \tag{H67}$$

with probability at least 0.99, where $\|X\|_{\mathrm{HS}} = \operatorname{tr}(X^2)$ is the Hilbert-Schmidt norm (Frobenius norm).

Proof. Let $\epsilon' = \epsilon/d$. We can employ union bound and Lemma H.2, H.3, H.4 to show that given $N_A = \mathcal{O}(1/(\epsilon')^2) = \mathcal{O}(d^2/\epsilon^2)$ and $N_B = \mathcal{O}(d^2 \log(d)/(\epsilon')^2) = \mathcal{O}(d^4 \log(d)/\epsilon^2)$ and $0 < \epsilon' < 0.1$, we have

$$|\hat{o}(\Phi_{\mathcal{N}}, P, Q) - \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})| < \epsilon' = \epsilon/d, \tag{H68}$$

with probability at least 0.99. Therefore, we have

$$\left\|\tilde{\Phi}_{\mathcal{N}}(N_A, N_B) - \Phi_{\mathcal{N}}\right\|_{\mathrm{HS}} = \sqrt{\frac{1}{d^2} \sum_{P,Q\in\{I,X,Y,Z\}^{\otimes n}} |\hat{o}(\Phi_{\mathcal{N}}, P, Q) - \operatorname{tr}((P\otimes Q)\Phi_{\mathcal{N}})|^2} < \epsilon \tag{H69}$$

with probability at least 0.99. This concludes the proof. □

I. Foundations for predicting extrinsic behaviors

Recall that Definition C.6 considers the predictability of properties in the world model. Here, we refer to a property as a function that maps a world model to a real value. Given action spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and an outcome space $\mathcal{B}$. There is a class of properties that are guaranteed to be predictable: the probability of an outcome $b \in \mathcal{B}$ for an experiment specified by the sequence of actions $x \in \mathcal{X}, y_1, \ldots, y_\ell \in \mathcal{Y}, z \in \mathcal{Z}$. We refer to such properties as the extrinsic behavior of the world model. Formally, this class of properties is given in the following definition.

Definition I.1 (Extrinsic behavior of world model)**.** Given action spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ and an outcome space $\mathcal{B}$. We consider extrinsic behavior to be a set of properties $\mathcal{F} = \{f : \mathcal{W} \to \mathbb{R}\}$. For any $f \in \mathcal{F}$, there is an experiment $E = (x \in \mathcal{X}, y_1 \in \mathcal{Y}, \ldots, y_\ell \in \mathcal{Y}, z \in \mathcal{Z})$, and an outcome $b \in \mathcal{B}$, such that

$$f(\mathcal{W}) = \mathrm{tr}(M_{zb}(\mathcal{E}_{y_\ell} \circ \ldots \circ \mathcal{E}_{y_1})(\rho_x)), \tag{I1}$$

for any world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$. We refer to f in the set as $f^{(\mathrm{behav})}_{x,y_1,\ldots,y_\ell,z,b}$.

In the following, we give several fundamental results regarding the task of predicting extrinsic behaviors. In particular, we will give rigorous guarantee for existing gate-set tomography protocols in terms of learning the extrinsic behavior of the world model.

I.1. Hardness in predicting extrinsic behaviors

The goal of this appendix is to prove the following theorem.

Theorem I.2 (Worst case complexity for predicting extrinsic behaviors; Restatement of Theorem 3)**.** *Consider finite sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ with $|\mathcal{B}| \geq 2$ and $|\mathcal{B}| = \mathcal{O}(\mathrm{poly}(|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|))$. Given $\epsilon, L > 0$. For any model class $\mathcal{Q}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$, there is an algorithm that uses*

$$\tilde{\mathcal{O}}\left(\frac{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}{\epsilon^2}\right) \tag{I2}$$

experiments to accurately predict $f^{(\mathrm{behav})}_{x,y_1,\ldots,y_L,z,b}(\mathcal{W})$ up to ϵ-error for all $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$ with high probability. Furthermore, there exists a model class $\mathcal{Q}$ over $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{B}$ such that

$$\Omega\left(\frac{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}{\epsilon^2}\right) \tag{I3}$$

experiments are required to accurately predict $f^{(\mathrm{behav})}_{x,y_1,\ldots,y_L,z,b}(\mathcal{W})$ up to ϵ-error for all $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$ with high probability.

The first part of the theorem is straightforward. We consider the algorithm that runs through all possible experiments with L CPTP maps. For an experiment specified by $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$, the algorithm performs each of the experiment for a number of

$$K = \mathcal{O}\left(\frac{\log\left(|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}||\mathcal{B}|\right)}{\epsilon^2}\right) \tag{I4}$$

times. For any world model $\mathcal{W} \in \mathcal{Q}$, using Hoeffding's inequality, the K experimental outcomes can be used to estimate the probability of an outcome $b \in \mathcal{B}$ for the experiment $E = (x, y_1, \ldots, y_L, z)$, i.e., $f^{(\mathrm{behav})}_{x,y_1,\ldots,y_L,z,b}(\mathcal{W})$, up to an error ϵ with probability $\geq 1 - 1/(10|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}||\mathcal{B}|)$. Using union bound, with probability ≥ 0.9, we can estimate $f^{(\mathrm{behav})}_{x,y_1,\ldots,y_L,z,b}(\mathcal{W})$ up to an error ϵ for all $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$. The total number of experiments used by the algorithm is

$$|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|K = \mathcal{O}\left(\frac{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|\log\left(|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}||\mathcal{B}|\right)}{\epsilon^2}\right) = \tilde{\mathcal{O}}\left(\frac{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}{\epsilon^2}\right), \tag{I5}$$

because of the assumption that $|\mathcal{B}| = \mathcal{O}(\text{poly}(|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{Z}|))$.

The second part of the theorem is more nontrivial. For this part of the proof, we will only utilize two elements in $\mathcal{B}$, which we denote as b_0, b_1. We begin by constructing a model class $\mathcal{Q}$. The model class $\mathcal{Q}$ contains a null world model $\mathcal{W}_0$ given by

$$\rho_x = |0\rangle\langle 0|, \qquad \forall x \in \mathcal{X}, \tag{I6}$$
$$\mathcal{E}_y(\rho) = |0\rangle\langle 0|, \qquad \forall y \in \mathcal{Y}, \tag{I7}$$
$$M_{zb_0} = \frac{1}{2} I, \qquad \forall z \in \mathcal{Z}, \tag{I8}$$
$$M_{zb_1} = \frac{1}{2} I, \qquad \forall z \in \mathcal{Z}, \tag{I9}$$
$$M_{zb} = 0, \qquad \forall z \in \mathcal{Z}, \forall b \in \mathcal{B} \setminus \{b_0, b_1\}, \tag{I10}$$

where I is the identity operator. The model class $\mathcal{Q}$ also contains another $2|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|$ world models $\mathcal{W}_{x,y_1,\ldots,y_L,z,s}$ indexed by $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, s = \pm 1$ satisfying

$$\rho_{x'} = |0\rangle\langle 0|, \qquad \forall x' \in \mathcal{X} \setminus \{x\}, \tag{I11}$$
$$\rho_x = |1\rangle\langle 1|, \tag{I12}$$
$$\mathcal{E}_{y_\ell}(|\ell\rangle\langle\ell|) = |\ell+1\rangle\langle\ell+1|, \qquad \forall 1 \le \ell \le L, \tag{I13}$$
$$\mathcal{E}_y(|\ell\rangle\langle\ell|) = |0\rangle\langle 0|, \qquad \forall 0 \le \ell \le L+1, y \in \mathcal{Y}, \text{ s.t.} \forall 1 \le \ell' \le L, y \ne y_{\ell'}, \tag{I14}$$
$$M_{zb_0} = \frac{1}{2}(I + 3s\epsilon|L+1\rangle\langle L+1|), \qquad \forall z \in \mathcal{Z}, \tag{I15}$$
$$M_{zb_1} = \frac{1}{2}(I - 3s\epsilon|L+1\rangle\langle L+1|), \qquad \forall z \in \mathcal{Z}, \tag{I16}$$
$$M_{zb} = 0, \qquad \forall z \in \mathcal{Z}, \forall b \in \mathcal{B} \setminus \{b_0, b_1\}, \tag{I17}$$

For the null world model $\mathcal{W}_0$, the outcome distribution is always a uniform distribution over b_0, b_1 for any experiments. For world model $\mathcal{W}_{x,y_1,\ldots,y_L,z,s}$, only for one experiment, i.e., when we consider $E = (x, y_1, \ldots, y_L, z)$, the outcome distribution is a biased distribution over b_0, b_1, in particular, we see b_0 with probability $\frac{1}{2} + \frac{3}{2}s\epsilon$ and b_1 with probability $\frac{1}{2} - \frac{3}{2}s\epsilon$. For all other experiments, the outcome distribution is again a uniform distribution over b_0, b_1. We will also denote $\mathcal{W}_{x,y_1,\ldots,y_L,z,s}$ as $\mathcal{W}_{E,s}$, where $E = (x, y_1, \ldots, y_L, z)$ is an experiment.

We consider the true world model to be $\mathcal{W}_0$ with probability $1/2$ and be $\mathcal{W}_{x,y_1,\ldots,y_L,z,s}$ for a particular choice of $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, s = \pm 1$ with probability $1/(4|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|)$. If there is a learning algorithm that could accurately predict $f^{(\text{behav})}_{x,y_1,\ldots,y_L,z,b}(\mathcal{W})$ up to ϵ-error for all $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$, then the algorithm can be used to check if the true world model is equal to $\mathcal{W}_0$. Recall that for $\mathcal{W}_0$, all experiments yield an outcome probability distribution that is uniform over $\{b_0, b_1\}$. For world model in $\mathcal{Q} \setminus \{\mathcal{W}_0\}$, the probability for one of the experiment specified $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$ will be a biased distribution. Hence, we can determine whether the true world model is equal to $\mathcal{W}_0$ by checking if for all $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$, the probability to obtain both b_0 and b_1 are close to $1/2$. This enables us to map the learning problem to a two-hypothesis testing problem. And the question becomes:

How many experiments are required to test if the true world model is equal to $\mathcal{W}_0$?

We will utilize the proof techniques given in [9, 30] to answer the above question.

A learning algorithm is represented by a tree where each node represents the collection of data received from all prior experiments. Each edge is label by an experiment E and an outcome $b \in \{b_0, b_1\}$ from that experiment. We only consider outcomes $\{b_0, b_1\}$ because the other outcomes happen with zero probability by construction. The probability for traversing that edge is the product of the probability for experiment E using the learning algorithm and the probability for seeing the outcome b for the experiment E. The depth T of the tree is the total number of experiments conducted by the learning algorithm. After conducting T experiments, we will arrive at a leaf node of the tree. The probability to arrive at a leaf node depends on the true world model.

We consider two events: when the true world model is $\mathcal{W}_0$, and when the true world model is not equal to $\mathcal{W}_0$. We distinguish the two events based on the data we collected from the T experiments, i,e., which leaf node has we arrived at. In order to successfully distinguish between the two events with a constant probability, we need the total variation distance of the probability distribution over the leaf nodes to be of $\Omega(1)$. Formally, this is known as LeCam's two point method.

Each leaf node ℓ is specified by the path from root to the leaf node, which is a sequence of T experiments and their corresponding outcomes. Hence, we write each leaf node as $\ell_{(E_1,\beta_1)\ldots,(E_T,\beta_T)}$, where $\beta_t \in \{b_0, b_1\} \subseteq \mathcal{B}$. For the null world model $\mathcal{W}_0$, the probability of a leaf node is given by

$$p_{\mathcal{W}_0}(\ell_{(E_1,\beta_1)\ldots,(E_T,\beta_T)}) = \prod_{t=1}^{T} \left(\frac{1}{2} p_{(E_1,\beta_1)\ldots,(E_{t-1},\beta_{t-1})}(E_t) \right), \tag{I18}$$

where $p_{(E_1,\beta_1)\ldots,(E_{t-1},\beta_{t-1})}(E_t)$ is the probability that the algorithm would perform the experiment E_t when the algorithm have run $t-1$ experiments $E_1, \ldots, E_{t-1}$ and the corresponding outcomes are $\beta_1, \ldots, \beta_{t-1} \in \{b_0, b_1\}$. For the alternative world model $\mathcal{W}_{E,s}$, the probability of a leaf node is

$$p_{\mathcal{W}_{E,s}}(\ell_{(E_1,\beta_1)\ldots,(E_T,\beta_T)}) = \prod_{t=1}^{T} \left(\frac{1}{2} \left(1 + 3s\epsilon\delta_{E_t,E} \operatorname{sign}(\beta_t)\right) p_{(E_1,\beta_1)\ldots,(E_{t-1},\beta_{t-1})}(E_t) \right), \tag{I19}$$

where $\operatorname{sign}(\beta_t) = 1$ if $\beta_t = b_0$ and $\operatorname{sign}(\beta_t) = -1$ if $\beta_t = b_1$. We can now write down the total variation distance between the probability distribution over the leaf nodes under the two events (true world model $= \mathcal{W}_0$ or true world model $\neq \mathcal{W}_0$),

$$\mathrm{TV} = \frac{1}{2} \sum_{\ell} \left| p_{\mathcal{W}_0}(\ell) - \mathop{\mathbb{E}}_{E,s} p_{\mathcal{W}_{E,s}}(\ell) \right| = \sum_{\ell:\, p_{\mathcal{W}_0}(\ell) \geq \mathbb{E}_{E,s}\, p_{\mathcal{W}_{E,s}}(\ell)} \left(p_{\mathcal{W}_0}(\ell) - \mathop{\mathbb{E}}_{E,s} p_{\mathcal{W}_{E,s}}(\ell) \right) \tag{I20}$$

$$\leq \sum_{\ell:\, p_{\mathcal{W}_0}(\ell) \geq p_{\mathcal{W}_{E,s}}(\ell)} p_{\mathcal{W}_0}(\ell)\eta \leq \eta, \tag{I21}$$

where the parameter η satisfies

$$\frac{\mathbb{E}_{E,s}\, p_{\mathcal{W}_{E,s}}(\ell)}{p_{\mathcal{W}_0}(\ell)} \geq 1 - \eta, \forall \text{ leaf node } \ell. \tag{I22}$$

We will now find a parameter η that satisfies the above condition. Given $\ell = \ell_{(E_1,\beta_1)\ldots,(E_T,\beta_T)}$, we have

$$\frac{\mathbb{E}_{E,s}\, p_{\mathcal{W}_{E,s}}(\ell)}{p_{\mathcal{W}_0}(\ell)} = \mathop{\mathbb{E}}_{E,s} \prod_{t=1}^{T} \left(1 + 3s\epsilon\delta_{E_t,E} \operatorname{sign}(\beta_t)\right) \tag{I23}$$

$$= \mathop{\mathbb{E}}_{E,s} \exp\left[\sum_{t=1}^{T} \log\left(1 + 3s\epsilon\delta_{E_t,E} \operatorname{sign}(\beta_t)\right) \right] \tag{I24}$$

$$\geq \exp\left[\mathop{\mathbb{E}}_{E,s} \sum_{t=1}^{T} \log\left(1 + 3s\epsilon\delta_{E_t,E} \operatorname{sign}(\beta_t)\right) \right] \tag{I25}$$

$$= \exp\left[\frac{1}{2} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{E} \log\left(1 - 9\epsilon^2\delta_{E_t,E}\right) \right] \tag{I26}$$

$$\geq \exp\left[\frac{1}{2} \sum_{t=1}^{T} \frac{1}{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|} \log\left(1 - 9\epsilon^2\right) \right] \tag{I27}$$

$$\geq \exp\left[-\sum_{t=1}^{T} \frac{1}{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|} 9\epsilon^2 \right] \tag{I28}$$

$$\geq 1 - \frac{9T\epsilon^2}{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}. \tag{I29}$$

The third line is Jensen's inequality. The fourth line uses the fact that $s = \pm 1$ uniformly. The fifth line uses the fact E is distributed uniformly over $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$. The second-to-last line uses $\log(1-x) \geq -2x, \forall x \in [0, 0.79]$ which is satisfied given $\epsilon < 0.29$. The last line uses $\exp(x) \geq 1 + x, \forall x \in \mathbb{R}$. Together, we can choose $\eta = \frac{9T\epsilon^2}{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}$ and we have established

$$\Omega(1) \leq \text{TV} \leq \frac{9T\epsilon^2}{|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|}. \tag{I30}$$

We have thus proved that the number of experiments must be at least $T = \Omega(|\mathcal{X}||\mathcal{Y}|^L|\mathcal{Z}|/\epsilon^2)$.

J. A general theorem for predicting extrinsic behaviors

In order to avoid the worst-case complexity proved in the previous subsection, we need to make stronger assumption about the true world model. We present one such assumption that is closely related to the assumption used in existing gate set tomography protocols [4, 7, 20, 39]. Intuitively, the assumption is that we can efficiently find a complete set of states that span the set of states we can generate using the world model, and a complete set of POVM elements where its span include the complete set of states. In the worst case, such as in the world models we constructed in the previous subsection, we cannot find a complete set of states and POVM elements efficiently. As a result, we see that the optimal complexity scales very badly (exponentially in L).

Before presenting the assumption, we will recall some basic concepts. Given a world model $\mathcal{W} = (\{\rho_x\}_{x\in\mathcal{X}}, \{\mathcal{E}_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}_z\}_{z\in\mathcal{Z}})$. We can compose a state ρ and a CPTP map $\mathcal{E}$ to prepare new state $\mathcal{E}(\rho)$. Similarly, we can compose a POVM $\mathcal{M}$ and a CPTP map $\mathcal{E}$ to compose a new POVM $\mathcal{M} \circ \mathcal{E}$ that is equivalent to first applying $\mathcal{E}$ then measure using $\mathcal{M}$. We refer to the states that could be prepared by composing a finite sequence of CPTP maps $\mathcal{E}_y$ on some initial state ρ_x as states that could be generated in $\mathcal{W}$.

J.1. Assumptions

We assume that we have found a linearly independent set of composed states $\{\rho_1, \ldots, \rho_{K_1}\}$ such that $\text{span}(\rho_1, \ldots, \rho_{K_1})$ contains all states that can be generated in $\mathcal{W}$, where span(...) consider all linear combinations with real coefficients. This is equivalent to stating that for any state ρ that can be generated in $\mathcal{W}$, there exists a unique set of coefficients $\alpha_1, \ldots, \alpha_{K_1} \in \mathbb{R}$ satisfying

$$\rho = \sum_{k_1=1}^{K_1} \alpha_{k_1} \rho_{k_1}. \tag{J1}$$

Because ρ and ρ_{k_1} for all $k_1 \in \{1, \ldots, K_1\}$ are quantum states, we have $\sum_{k_1} \alpha_{k_1} = 1$. We consider $R_1 > 0$ to be the constant such that

$$\sum_{k_1=1}^{K_1} |\alpha_{k_1}| \leq R_1 \tag{J2}$$

for any quantum state ρ in $\text{span}(\rho_1, \ldots, \rho_{K_1})$.

Remark 4. $R_1 < \infty$ because a finite-dimensional quantum state space is compact.

We also assume that we have found a set of composed POVM elements $\{M_1, \ldots, M_{K_2}\}$ such that $\text{span}(\rho_1, \ldots, \rho_{K_1}) \subseteq \text{span}(M_1, \ldots, M_{K_2})$. This assumption implies that for any $\alpha, \alpha' \in \mathbb{R}^{K_1}$, if

$$\forall k_2 \in \{1, \ldots, K_2\}, \text{tr}\left(M_{k_2} \sum_{k_1=1}^{K_1} \alpha_{k_1} \rho_{k_1}\right) = \text{tr}\left(M_{k_2} \sum_{k_1=1}^{K_1} \alpha'_{k_1} \rho_{k_1}\right), \tag{J3}$$

then $\alpha = \alpha'$. In particular, we consider $R_2 > 0$ to be the constant such that

$$\left\|\alpha - \alpha'\right\|_1 \leq R_2 \sum_{k_2=1}^{K_2} \left| \mathrm{tr}\left(M_{k_2} \sum_{k_1=1}^{K_1} \alpha_{k_1} \rho_{k_1} \right) - \mathrm{tr}\left(M_{k_2} \sum_{k_1=1}^{K_1} \alpha'_{k_1} \rho_{k_1} \right) \right| \tag{J4}$$

for any $\alpha, \alpha' \in \mathbb{R}^{K_1}$.

Remark 5. $R_2 < \infty$ because the homogeneity of Eq. (J4) enables maximization to find R_2 over the compact space $\|\alpha - \alpha'\|_1 = 1$.

J.2. Learning a frame

We are now ready to present an efficient learning algorithm that can be used to predict the extrinsic behavior of the world model $\mathcal{W}$. For each $k_2 = 1, \ldots, K_2$, we consider the POVM element M_{k_2} to be a unit vector in K_2-dimensional vector space,

$$M_{k_2} \to \hat{e}_{k_2}. \tag{J5}$$

For each $k_1 = 1, \ldots, K_1$, we consider the state ρ_{k_1} to be a vector $w_{k_1} \in \mathbb{R}^{K_2}$,

$$\rho_{k_1} \to w_{k_1} \equiv \sum_{k_2=1}^{K_2} w_{k_1,k_2} \hat{e}_{k_2}, \tag{J6}$$

where w_{k_1,k_2} is an estimate for $\mathrm{tr}(M_{k_2}\rho_{k_1})$, such that

$$\Pr\left[\ |w_{k_1,k_2} - \mathrm{tr}(M_{k_2}\rho_{k_1})| > \tilde{\epsilon}_w\ \right] \leq \delta. \tag{J7}$$

We can obtain an estimate w_{k_1,k_2} with the above guarantee using $\mathcal{O}(\log(1/\delta)/\tilde{\epsilon}_w^2)$ experiments. We simply prepare ρ_{k_1} and measure the POVM associated to M_{k_2}, then compute the average of the indicator function on whether the POVM element M_{k_2} is the outcome. Because indicator function of an event is a bounded random variable, Hoeffding's inequality gives us the above rigorous guarantee.

J.3. Learning states

For the initial states ρ_x with $x \in \mathcal{X}$, we represent ρ_x as a K_1-dimensional real vector

$$\rho_x \to v_x \equiv \sum_{k_1=1}^{K_1} v_{x,k_1} \hat{e}_{k_1}, \tag{J8}$$

where v_x is an optimum of the following optimization problem

$$\mathrm{OPT}(v_x) = \min_{\substack{\alpha \in \mathbb{R}^{K_1}, \|\alpha\|_1 \leq R_1, \\ \sum_{k_1=1}^{K_1} \alpha_{k_1} = 1}} \left\| \sum_{k_1=1}^{K_1} \alpha_{k_1} w_{k_1} - \sum_{k_2=1}^{K_2} v'_{x,k_2} \hat{e}_{k_2} \right\|_1, \tag{J9}$$

and v'_{x,k_2} is an estimate for $\mathrm{tr}(M_{k_2}\rho_x)$, such that

$$\Pr\left[\ |v'_{x,k_2} - \mathrm{tr}(M_{k_2}\rho_x)| > \tilde{\epsilon}_v\ \right] \leq \delta. \tag{J10}$$

Because the above optimization is a convex optimization, one could solve for α_x efficiently. The purpose of the optimization problem is to project the vector $\sum_{k_2=1}^{K_2} v'_{x,k_2} \hat{e}_{k_2}$ onto the space spanned by $v_1, \ldots, v_{K_1}$. For a fixed $\tilde{\epsilon}_v, \delta > 0$, the total number of experiments for learning representations of the states is $\mathcal{O}(K_2|\mathcal{X}|\log(1/\delta)/\tilde{\epsilon}_v^2)$.

J.4. Learning CPTP maps

For the CPTP maps $\mathcal{E}_y$ with $y \in \mathcal{Y}$, we represent $\mathcal{E}_y$ as a $K_1 \times K_1$ real matrix of the form

$$\mathcal{E}_y \to A_y \equiv \sum_{k_1=1}^{K_1} \sum_{k_1'=1}^{K_1} A_{y,k_1,k_1'} \hat{e}_{k_1} \hat{e}_{k_1'}^T, \tag{J11}$$

where for each fixed value of k_1', the K_1-dimensional vector $A_{y,(\cdot),k_1'}$ is an optimum of the following convex optimization problem

$$\mathrm{OPT}(A_{y,(\cdot),k_1'}) = \min_{\substack{\alpha \in \mathbb{R}^{K_1}, \|\alpha\|_1 \le R_1, \\ \sum_{k_1=1}^{K_1} \alpha_{k_1} = 1}} \left\| \sum_{k_1=1}^{K_1} \alpha_{k_1} w_{k_1} - \sum_{k_2=1}^{K_2} A'_{y,k_2,k_1'} \hat{e}_{k_2} \right\|_1, \tag{J12}$$

and $A'_{y,k_2,k_1'}$ is an estimate for $\mathrm{tr}(M_{k_2} \mathcal{E}_y(\rho_{k_1'}))$, such that

$$\Pr\left[\, |A'_{y,k_2,k_1'} - \mathrm{tr}(M_{k_2} \mathcal{E}_y(\rho_{k_1'}))| > \tilde{\epsilon}_A \, \right] \le \delta. \tag{J13}$$

Similar to before, the purpose of the optimization is to project the vector $\sum_{k_2=1}^{K_2} A'_{y,k_2,k_1'} \hat{e}_{k_2}$ to the space formed by $v_1, \ldots, v_{K_1}$. For a fixed $\tilde{\epsilon}_A, \delta > 0$, the total number of experiments for learning representations of the CPTP maps is $\mathcal{O}(K_1 K_2 |\mathcal{Y}| \log(1/\delta)/\tilde{\epsilon}_A^2)$.

J.5. Learning POVMs

For the POVM elements M_{zb} with $z \in \mathcal{Z}, b \in \mathcal{B}$, we represent M_{zb} as a K_1-dimensional real vector

$$M_{zb} \to u_{zb} \equiv \sum_{k_1=1}^{K_1} u_{z,b,k_1} \hat{e}_{k_1}, \tag{J14}$$

where u_{z,b,k_1} is an estimate for $\mathrm{tr}(M_{zb} \rho_{k_1})$, such that

$$\Pr\left[\, |u_{z,b,k_1} - \mathrm{tr}(M_{zb} \rho_{k_1})| > \tilde{\epsilon}_u \, \right] \le \delta. \tag{J15}$$

For a fixed $\tilde{\epsilon}_u, \delta > 0$, the total number of experiments for learning representations of the POVMs is $\mathcal{O}(K_1 |\mathcal{Z}| \log(1/\delta)/\tilde{\epsilon}_u^2)$. The reason that we don't need an extra factor of $|\mathcal{B}|$ is because when we measure the POVM $\mathcal{M}_z$, we can simultaneously estimate c_{z,b,k_1} for all $b \in \mathcal{B}$.

J.6. Prediction procedure and rigorous guarantee

During the prediction phase, we predict the probability for obtaining an outcome $b \in \mathcal{B}$ after running the experiment $E = (x, y_1, \ldots, y_L, z)$ to be

$$u_{zb}^T P A_{y_L} P \ldots P A_{y_1} v_x, \tag{J16}$$

where P is the projection to the convex set $\{\alpha \in \mathbb{R}^{K_1} | \sum_{k_1} \alpha_{k_1} = 1, \|\alpha\|_1 \le R_1\}$. We now give a rigorous performance guarantee for this algorithm.

Theorem J.1 (Predicting extrinsic behaviors; Restatement of Theorem 4)**.** *Assume that we have found a complete set of linearly independent states and POVMs. Using the proposed algorithm, we can predict* $\mathrm{tr}(M_{zb}(\mathcal{E}_{y_L} \circ \ldots \circ \mathcal{E}_{y_1})(\rho_x))$ *to* ϵ *error for all* $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}, b \in \mathcal{B}$ *using a total of*

$$\tilde{\mathcal{O}}\left(\frac{|\mathcal{X}| + L^2 |\mathcal{Y}| + |\mathcal{Z}|}{\epsilon^2} \right) \tag{J17}$$

experiments, where $\tilde{\mathcal{O}}(\cdot)$ *neglects logarithmic factors and considers* K_1, K_2, R_1, R_2 *to be constant.*

J.7. Proof of Theorem J.1 — Step A. Representations of target outputs

From Eq. (J1) and (J2), $\forall x \in \mathcal{X}$, we can write ρ_x as

$$\rho_x = \sum_{k_1=1}^{K_1} v^*_{x,k_1} \rho_{k_1}, \tag{J18}$$

for some K_1-dimensional vector v^*_x with $\|v^*_x\|_1 \le R_1$. Similarly, Eq. (J1) and (J2) shows that for any $k'_1 \in \{1, \dots, K_1\}$, we can write $\mathcal{E}_y(\rho_{k'_1})$ as

$$\mathcal{E}_y(\rho_{k'_1}) = \sum_{k_1=1}^{K_1} A^*_{y,k_1,k'_1} \rho_{k_1}, \tag{J19}$$

for some K_1-dimensional vector $A^*_{y,(\cdot),k'_1}$ with $\left\|A^*_{y,(\cdot),k'_1}\right\|_1 \le R_1$. Using this representation, for an experiment specified by $x \in \mathcal{X}, y_1, \dots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$, the probability to obtain the measurement outcome $b \in \mathcal{B}$ can be written as

$$\operatorname{tr}(M_{zb}(\mathcal{E}_{y_L} \circ \dots \circ \mathcal{E}_{y_1})(\rho_x)) = \sum_{k_1=1}^{K_1} (A^*_{y_L} \dots A^*_{y_1} v^*_x)_{k_1} \operatorname{tr}(M_{zb}\rho_{k_1}) = (u^*_{zb})^T A^*_{y_L} \dots A^*_{y_1} v^*_x, \tag{J20}$$

where we let $u^*_{z,b,k_1} \equiv \operatorname{tr}(M_{zb}\rho_{k_1})$ and $u^*_{zb} = \sum_{k_1} u^*_{z,b,k_1} \hat{e}_{k_1}$.

J.8. Proof of Theorem J.1 — Step B. Error analysis for the learned representations

We begin by comparing the two vectors v^*_x and v_x, where v_x is the representation learned from experiments; see Eq. (J8). We can bound the difference as follows,

$$\begin{aligned}
\|v^*_x - v_x\|_1 &\le R_2 \sum_{k_2=1}^{K_2} \left| \operatorname{tr}(M_{k_2}\rho_x) - \sum_{k_1=1}^{K_1} v_{x,k_1} \operatorname{tr}(M_{k_2}\rho_{k_1}) \right| && \text{(J21)}\\
&\le R_2 \sum_{k_2=1}^{K_2} \left(\left| v'_{x,k_2} - \sum_{k_1=1}^{K_1} v_{x,k_1} w_{k1,k2} \right| + \tilde{\epsilon}_v + \|v_x\|_1 \tilde{\epsilon}_w \right) && \text{(J22)}\\
&= R_2 \mathrm{OPT}(v_x) + R_2 K_2 \left(\tilde{\epsilon}_v + \|v_x\|_1 \tilde{\epsilon}_w\right), && \text{(J23)}\\
&\le R_2 K_2 \left(\tilde{\epsilon}_v + \|v^*_x\|_1 \tilde{\epsilon}_w\right) + R_2 K_2 \left(\tilde{\epsilon}_v + \|v_x\|_1 \tilde{\epsilon}_w\right), && \text{(J24)}\\
&\le 2R_2 K_2 \left(\tilde{\epsilon}_v + R_2 \tilde{\epsilon}_w\right), && \text{(J25)}
\end{aligned}$$

with probability at least $1 - (K_2 + 1)\delta$. The first line follows from Eq. (J4) and (J18). The second line follows from Eq. (J7), (J10), and union bound. The third line follows from Eq. (J9). The fourth line follows from considering $\alpha = v^*_x$ in the optimization problem given at Eq. (J9), and utilize the following bounds,

$$\begin{aligned}
&\left\| \sum_{k_1=1}^{K_1} v^*_{x,k_1} w_{k_1} - \sum_{k_2=1}^{K_2} v'_{x,k_2} \hat{e}_{k_2} \right\|_1 && \text{(J26)}\\
&\le \sum_{k_2=1}^{K_2} \left| \sum_{k_1=1}^{K_1} v^*_{x,k_1} \operatorname{tr}(M_{k_2}\rho_{k_1}) - \operatorname{tr}(M_{k_2}\rho_x) \right| + K_2 \left(\tilde{\epsilon}_v + \|v^*_x\|_1 \tilde{\epsilon}_w\right) && \text{(J27)}\\
&= \sum_{k_2=1}^{K_2} |\operatorname{tr}(M_{k_2}\rho_x) - \operatorname{tr}(M_{k_2}\rho_x)| + K_2 \left(\tilde{\epsilon}_v + \|v^*_x\|_1 \tilde{\epsilon}_w\right) = K_2 \left(\tilde{\epsilon}_v + \|v^*_x\|_1 \tilde{\epsilon}_w\right). && \text{(J28)}
\end{aligned}$$

The last line follows from $\|v_x^*\|_1, \|v_x\|_1 \leq R_1$.

We can compare $A^*_{y,k_1,k_1'}$ and $A_{y,k_1,k_1'}$ by employing the same analysis and replacing Eq. (J18) with (J19), Eq. (J8) with (J11), Eq. (J10) with (J13), and Eq. (J9) with (J12). For all $y \in \mathcal{Y}, k_1' \in \{1, \ldots, K_1\}$, the analysis shows that the following,

$$\left\| A^*_{y,(\cdot),k_1'} - A_{y,(\cdot),k_1'} \right\|_1 \leq 2R_2K_2 \left(\tilde{\epsilon}_A + R_2\tilde{\epsilon}_w\right), \tag{J29}$$

happens with probability at least $1-(K_2+1)\delta$. And recall from Eq. (J15), we have $|u_{z,b,k_1} - u^*_{z,b,k_1}| \leq \tilde{\epsilon}_u$ with probability at least $1-\delta$. Together, with probability at least $1 - |\mathcal{X}|(K_2+1)\delta - |\mathcal{Y}|K_1(K_2+1)\delta - |\mathcal{Z}||\mathcal{B}|\delta = 1 - ((|\mathcal{X}| + K_1|\mathcal{Y}|)(K_2+1) + |\mathcal{Z}||\mathcal{B}|)\delta$, we have

$$\sum_{k_1=1}^{K_1} \left| v^*_{x,k_1} - v_{x,k_1} \right| \leq 2R_2K_2 \left(\tilde{\epsilon}_v + R_2\tilde{\epsilon}_w\right), \qquad \forall x \in \mathcal{X}, \tag{J30}$$

$$\sum_{k_1=1}^{K_1} \left| A^*_{y,k_1,k_1'} - A_{y,k_1,k_1'} \right| \leq 2R_2K_2 \left(\tilde{\epsilon}_A + R_2\tilde{\epsilon}_w\right), \qquad \forall y \in \mathcal{Y}, \forall k_1' \in \{1, \ldots, K_1\}, \tag{J31}$$

$$\left| u^*_{z,b,k_1} - u_{z,b,k_1} \right| \leq \tilde{\epsilon}_u, \qquad \forall z \in \mathcal{Z}, b \in \mathcal{B}, \forall k_1 \in \{1, \ldots, K_1\}. \tag{J32}$$

This provides a set of error bounds for the learned representations.

J.9. Proof of Theorem J.1 — Step C. Error analysis for the prediction

We will now analyze the difference between the prediction and the true answer. We first define a linear function mapping a K_1-dimensional vector α to a matrix

$$\rho(\alpha) = \sum_{k_1=1}^{K_1} \alpha_{k_1} \rho_{k_1}. \tag{J33}$$

Because ρ_{k_1} is a quantum state, we have

$$\|\rho(\alpha)\|_1 \leq \sum_{k_1=1}^{K_1} |\alpha_{k_1}| \, \|\rho_{k_1}\|_1 = \|\alpha\|_1 . \tag{J34}$$

Note that $\|\rho(\alpha)\|_1$ is the trace norm for a matrix, while $\|\alpha\|_1$ is a vector one-norm. For an experiment specified by $x \in \mathcal{X}, y_1, \ldots, y_L \in \mathcal{Y}, z \in \mathcal{Z}$, the difference between the probability for obtaining $b \in \mathcal{B}$ and the predicted probability is

$$\left| u_{zb}^T P A_{y_L} P \ldots P A_{y_1} v_x - (u^*_{zb})^T A^*_{y_L} \ldots A^*_{y_1} v^*_x \right| \tag{J35}$$

$$\leq \max_{k_1} \left| u_{z,b,k_1} - u^*_{z,b,k_1} \right| \left\| P A_{y_L} P \ldots P A_{y_1} v_x \right\|_1 \tag{J36}$$

$$+ \left| \operatorname{tr}\left(M_{zb} \rho(P A_{y_L} P \ldots P A_{y_1} v_x)\right) - \operatorname{tr}\left(M_{zb} \rho(A^*_{y_L} \ldots A^*_{y_1} v^*_x)\right) \right| \tag{J37}$$

$$\leq \tilde{\epsilon}_u R_1 + \left\| \rho(P A_{y_L} P \ldots P A_{y_1} v_x) - \rho(A^*_{y_L} \ldots A^*_{y_1} v^*_x) \right\|_1 . \tag{J38}$$

The first inequality is a telescoping sum with a triangle inequality. The second inequality follows from Eq. (J32) and the fact that $\|M_{zb}\|_\infty \leq 1$.

We will now analyze the second term in the above equation. We will prove that

$$\left\| \rho(P A_{y_L} P \ldots P A_{y_1} v_x) - \rho(A^*_{y_L} \ldots A^*_{y_1} v^*_x) \right\|_1 \tag{J39}$$

$$\leq 2R_2K_2 \left(\tilde{\epsilon}_v + R_2\tilde{\epsilon}_w\right) + 2LR_1R_2K_2 \left(\tilde{\epsilon}_A + R_2\tilde{\epsilon}_w\right) \tag{J40}$$

by induction on L. For the base case $L = 0$, from Eq. (J30) and Eq. (J34), we see that

$$\|\rho(v_x) - \rho(v^*_x)\|_1 \leq \|v_x - v^*_x\|_1 \leq 2R_2K_2 \left(\tilde{\epsilon}_v + R_2\tilde{\epsilon}_w\right). \tag{J41}$$

Suppose that the claim holds for $L-1$. Then

$$\left\|\rho(PA_{y_L}P\dots PA_{y_1}v_x)-\rho(A^*_{y_L}\dots A^*_{y_1}v^*_x)\right\|_1 \tag{J42}$$
$$\le \left\|\rho(PA_{y_L}(PA_{y_{L-1}}\dots PA_{y_1}v_x))-\rho(A^*_{y_L}(PA_{y_{L-1}}\dots PA_{y_1}v_x))\right\|_1 \tag{J43}$$
$$+\left\|\rho(A^*_{y_L}(PA_{y_{L-1}}\dots PA_{y_1}v_x))-\rho(A^*_{y_L}(A^*_{y_{L-1}}\dots A^*_{y_1}v^*_x))\right\|_1 \tag{J44}$$
$$\le \left\|PA_{y_L}(PA_{y_{L-1}}\dots PA_{y_1}v_x)-A^*_{y_L}(PA_{y_{L-1}}\dots PA_{y_1}v_x)\right\|_1 \tag{J45}$$
$$+\left\|\mathcal{E}_{y_L}\left(\rho(PA_{y_{L-1}}\dots PA_{y_1}v_x-A^*_{y_{L-1}}\dots A^*_{y_1}v^*_x)\right)\right\|_1 \tag{J46}$$
$$\le \sum_{k_1=1}^{K_1}\left(\max_{k'_1}\left|PA_{y_L,k_1,k'_1}-A^*_{y_L,k_1,k'_1}\right|\sum_{k'_1=1}^{K_1}\left|(PA_{y_{L-1}}\dots PA_{y_1}v_x)_{k'_1}\right|\right) \tag{J47}$$
$$+\left\|\rho(PA_{y_{L-1}}\dots PA_{y_1}v_x-A^*_{y_{L-1}}\dots A^*_{y_1}v^*_x)\right\|_1 \tag{J48}$$
$$\le 2R_1R_2K_2\left(\tilde\epsilon_A+R_2\tilde\epsilon_w\right) \tag{J49}$$
$$+2R_2K_2\left(\tilde\epsilon_v+R_2\tilde\epsilon_w\right)+2(L-1)R_1R_2K_2\left(\tilde\epsilon_A+R_2\tilde\epsilon_w\right) \tag{J50}$$
$$=2R_2K_2\left(\tilde\epsilon_v+R_2\tilde\epsilon_w\right)+2LR_1R_2K_2\left(\tilde\epsilon_A+R_2\tilde\epsilon_w\right). \tag{J51}$$

The first inequality follows from triangle inequality and the linearity of $\rho(\alpha)$. The second inequality follows from the action of $\mathcal{E}_y$ given in Eq. (J19). The third inequality uses two basic inequalities: $\|(A-A')x\|\le\sum_i(\max_j|(A-A')_{ij}|\sum_j x_j),\forall A,A'\in\mathbb{R}^{k\times k},x\in\mathbb{R}^k$, and $\|\mathcal{E}(X)\|_1\le\|X\|_1$ for CPTP map $\mathcal{E}$ and Hermitian matrix X. The fourth inequality uses induction hypothesis, Eq. (J31), and the fact that $\sum_k|(Px)_k|\le R_1$ for any $x\in\mathbb{R}^{K_1}$.

J.10. Proof of Theorem J.1 — Step D. Putting everything together

Together, we consider the following parameter choices,

$$\frac{1}{\delta}=100((|\mathcal{X}|+K_1|\mathcal{Y}|)(K_2+1)+|\mathcal{Z}||\mathcal{B}|), \tag{J52}$$
$$\frac{1}{\tilde\epsilon_w}=\frac{16LR_1R_2^2K_2}{\epsilon}, \tag{J53}$$
$$\frac{1}{\tilde\epsilon_v}=\frac{8R_2K_2}{\epsilon}, \tag{J54}$$
$$\frac{1}{\tilde\epsilon_A}=\frac{8LR_1R_2K_2}{\epsilon}, \tag{J55}$$
$$\frac{1}{\tilde\epsilon_u}=\frac{4R_1}{\epsilon}, \tag{J56}$$

to ensure that with probability at least 0.99, for any experiment specified by $x\in\mathcal{X},y_1,\dots,y_L\in\mathcal{Y},z\in\mathcal{Z}$, the difference between the actual probability for obtaining the measurement outcome $b\in\mathcal{B}$ and the predicted probability is bounded above by ϵ,

$$\left|u_{zb}^TPA_{y_L}P\dots PA_{y_1}v_x-\mathrm{tr}(M_{zb}(\mathcal{E}_{y_L}\circ\dots\circ\mathcal{E}_{y_1})(\rho_x))\right|\le\epsilon. \tag{J57}$$

By aggregating the number of experiments for learning a frame, states, maps, and POVMs, the total number of experiment is of order

$$\frac{L^2R_1^2R_2^4K_2^2+|\mathcal{X}|R_2^2K_2^3+L^2|\mathcal{Y}|K_1K_2^3R_1^2R_2^2+|\mathcal{Z}|R_1^2K_1}{\epsilon^2}\log\left(K_2|\mathcal{X}|+K_1K_2|\mathcal{Y}|+|\mathcal{Z}||\mathcal{B}|\right) \tag{J58}$$
$$=\tilde{\mathcal{O}}\left(\frac{|\mathcal{X}|+L^2|\mathcal{Y}|+|\mathcal{Z}|}{\epsilon^2}\right), \tag{J59}$$

where $\tilde{\mathcal{O}}(\cdot)$ neglects logarithmic contributions and considers K_1,K_2,R_1,R_2 to be constant.

K. Related work

In this appendix, we present the connection of the theory developed in this work and existing works. We will refer to basic concepts developed in some of the previous appendices. In particular, Appendix A on the definition of world models, Appendix C on the definition of model classes and learning intrinsic descriptions, and Appendix I on the definition of predicting extrinsic behaviors.

K.1. Gate set tomography

The most relevant literature to the theory of world models developed in this work is gate set tomography. Here we provide a brief review on gate set tomography. We refer the readers to two comprehensive reviews on gate set tomography [20, 39]. Ref. [20] gives the basic concepts and [39] provides both the fundamental ideas and recent progress on gate set tomography. An experimental demonstration of gate set tomography is given in [4]. An open-sourced software for gate set tomography has been developed in [15].

The main goal of gate set tomography (GST) is to characterize how quantum processes and logical gates affect the qubits in the device. This is closely related to the task of quantum process tomography (QPT). However, in quantum process tomography, one assumes perfect state preparations and perfect POVM measurements. The key differences between GST and QPT are: (1) the lack of assumption on perfect states and measurements; (2) the need to benchmark multiple quantum processes (gates) at once. Because GST does not assume the ability to prepare perfect states and measurements, existing protocols learn the relation between different gates instead of the intrinsic physical description of each gate. This problem is referred to as gauge freedom in GST.

When we vectorize state and POVM element as $\mathrm{vec}(\rho_x)$ and $\mathrm{vec}(M_{zb})$ and write the CPTP maps $\mathcal{E}_y$ as a matrix A_y, then we have

$$\mathrm{tr}(M_{zb}\mathcal{E}_y(\rho_x)) = \mathrm{vec}(M_{zb})^T A_y \mathrm{vec}(\rho_x) = (M^{-1}\mathrm{vec}(M_{zb}))^T (M A_y M^{-1})(M\mathrm{vec}(\rho_x)). \tag{K1}$$

The following transformation is known as a gauge transformation

$$\mathrm{vec}(\rho_x) \to M\mathrm{vec}(\rho_x), \tag{K2}$$

$$A_y \to M A_y M^{-1}, \tag{K3}$$

$$\mathrm{vec}(M_{zb}) \to M^{-1}\mathrm{vec}(M_{zb}), \tag{K4}$$

which is parameterized by an invertible matrix M. Most existing GST protocols are designed to learn the vectorization and matricization up the the gauge freedom, specified by the matrix M. Due to the gauge M, it has been difficult to provide a fully rigorous statistical analysis of GST. The key problem is that the gauge freedom makes it challenging to define errors in GST. And we are not aware of a rigorous proof showing the required number of experiments to perform GST up to certain error.

We begin with a summary of the difference between the theory on learning world models developed in this work and the existing literature on gate set tomography [39]. According to the review [39], GST is tomography of a novel entity, which is not the individual description of each gate, but some form of relations between the gates. Based on our theory, we can formalize this novel entity as a non-physical model capable of predicting the extrinsic behaviors of the quantum device under the following assumption: one can (efficiently) find a complete basis of states and POVMs by composing existing states, gates, and measurements. Without this assumption, Theorem 3 shows that learning the extrinsic behavior can be extremely inefficient. And under this assumption, Theorem 4 provides a rigorous algorithm, that shares many common aspects with GST, for predicting extrinsic behaviors that improves significantly upon the worst case complexity given in Theorem 3.

In contrast, learning intrinsic physical descriptions of the operations in the device is significantly more challenging than performing GST. In certain scenarios, it is impossible to learn the intrinsic descriptions to arbitrary accuracy as shown in Theorem 6. However, we show that it is possible in

many natural scenarios. We can learn the intrinsic description up to any accuracy when the world model has an unknown pure state and a universal set of unknown gates (Theorem 2). Hence, being "informationally complete" for GST is easier than being "informationally complete" for learning intrinsic descriptions.

We now present several questions that illustrate the difference between the theory on learning world models developed in this work and gate set tomography.

a. Is gauge transformation the same as weakly indistinguishable? No. When two world models are related by the gauge transformation defined in GST, it is true that the two world models are weakly indistinguishable, i.e., no experiments can distinguish the two world models. However, there are pairs of weakly indistinguishable world models that are not related by a gauge transformation. One such example has been presented in the main text. Consider the following two distinct physical realities in a single-qubit system,

$$\mathcal{W}^A_{\mathrm{HT}}: \quad \rho^A_0 = I/2, \qquad \mathcal{E}^A_H(\rho) = H\rho H^\dagger, \qquad \mathcal{E}^A_T(\rho) = T\rho T^\dagger, \qquad \mathcal{M}^A_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{K5}$$

$$\mathcal{W}^B_{\mathrm{HT}}: \quad \rho^B_0 = I/2, \qquad \mathcal{E}^B_H(\rho) = I/2, \qquad \mathcal{E}^B_T(\rho) = I/2, \qquad \mathcal{M}^B_0 = \{|0\rangle\langle 0|, |1\rangle\langle 1|\}, \tag{K6}$$

with actions $\mathcal{X} = \{0\}, \mathcal{Y} = \{H, T\}, \mathcal{Z} = \{0\}$. Because $\mathcal{E}^A_H \neq \mathcal{E}^A_T$ but $\mathcal{E}^B_H = \mathcal{E}^B_T$, it is impossible that the two world models are related by a gauge transformation. However, as we have discussed in the main text, the two world models cannot be distinguished by any experiment, hence they are weakly indistinguishable. Together, we see that gauge equivalence as defined in GST implies weakly indistinguishability, but two devices that are weakly indistinguishable does not imply gauge equivalence in GST. This means the set of relations defined by gauge equivalence in GST is a subset of the relations defined by weakly indistinguishable world models.

b. Can we always learn a world model up to gauge transformation? No. While GST removes the assumption on perfect state preparations and measurements, GST still require assumptions to work. It is not true that we can always learn a world model up to a gauge transformation using existing GST protocols. This is already evident in the above example. The two world models $\mathcal{W}^A_{\mathrm{HT}}, \mathcal{W}^B_{\mathrm{HT}}$ are not related by a gauge transformation, but they are weakly indistinguishable. Hence by Proposition E.2, we cannot learn to distinguish the two world models from experiments.

c. Is gauge transformation the same as strongly indistinguishable (equivalence)? No. As Theorem A.8 has shown, equivalent world models are related by unitary or anti-unitary transformation. But two world models related by the gauge transformation defined in GST may not be related by unitary or anti-unitary transformation. For example, consider two d-dimensional world models over the action spaces $\mathcal{X} = \{\sigma\}_{\sigma:\mathrm{state}}, \mathcal{Y} = \{U\}_{U\in\mathrm{SU}(d)}, \mathcal{Z} = \{0\}$ and outcome space $\mathcal{B} = \{1, \ldots, d\}$: $\mathcal{W}^A = \left(\{\rho^A_x\}_{x\in\mathcal{X}}, \{\mathcal{E}^A_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}^A_z\}_{z\in\mathcal{Z}}\right)$, where

$$\rho^A_\sigma = (1-\epsilon)\sigma + \epsilon\frac{I}{d}, \qquad \forall\sigma : \mathrm{state}, \tag{K7}$$

$$\mathcal{E}^A_U(\rho) = U\rho U^\dagger, \qquad \forall U \in \mathrm{SU}(d), \tag{K8}$$

$$\mathcal{M}^A_0 = \{|b\rangle\langle b|\}_{b=1,\ldots,d}, \tag{K9}$$

and $\mathcal{W}^B = \left(\{\rho^B_x\}_{x\in\mathcal{X}}, \{\mathcal{E}^B_y\}_{y\in\mathcal{Y}}, \{\mathcal{M}^B_z\}_{z\in\mathcal{Z}}\right)$, where

$$\rho^B_\sigma = \sigma, \qquad \forall\sigma : \mathrm{state}, \tag{K10}$$

$$\mathcal{E}^B_U(\rho) = U\rho U^\dagger, \qquad \forall U \in \mathrm{SU}(d), \tag{K11}$$

$$\mathcal{M}^B_0 = \left\{(1-\epsilon)|b\rangle\langle b| + \epsilon\frac{I}{d}\right\}_{b=1,\ldots,d}, \tag{K12}$$

We can see that $\mathcal{W}^A$ has depolarized initial states, while $\mathcal{W}^B$ has depolarized measurements. The two world models are not related by a unitary or anti-unitary transformation, hence they describe distinct physical realities. However, it is not hard to show that the two world models are related by a gauge transformation M.

d. Are two gauge-equivalent devices always physically the same? No. We can have two noisy devices with depolarization noise happening in the states or in the measurements; see the above example of $\mathcal{W}^A$ and $\mathcal{W}^B$. The two devices are gauge equivalent. But these two noise processes are not physically the same. And hence the two devices are not physically the same. The inability to distinguish the two world models $\mathcal{W}^A$ and $\mathcal{W}^B$ is due to the lack of sufficiently information actions. This is similar to the case when one performs quantum state tomography with computational basis measurements. A coherent $|+\rangle\langle+|$ state is indistinguishable from a maximally mixed state $I/2$ under computational basis measurements. However, the two states are not physically the same. The indistinguishability arises not from being physically the same, but from the lack of useful actions.

e. How does gauge transformation defined in GST relate to our theory on world models? Suppose that we have a d-dimensional world model $\mathcal{W}$, where we are assumed that we can identify a set of (composed) states and a set of (composed) POVM elements, such that both sets span the entire d-dimensional quantum state space. Then every world model $\mathcal{W}'$ that is weakly indistinguishable from $\mathcal{W}$ is related by a gauge transformation. Hence, under the assumption, we can equate weakly indistinguishability and gauge equivalence in GST. This assumption is critical and has been explicitly or implicitly assumed in existing GST protocols [39]. A similar assumption is made in Theorem 4 (restated in Theorem J.1) for developing an efficient algorithm for learning extrinsic behaviors.

f. What entity is GST learning according to our theory of world models? In this in-depth review on GST [39], the authors noted multiple times that GST protocols are learning the relations between the gates instead of the descriptions for individual gates. The description found on each gate may not be endowed with a physical meaning — for some gauge M, the resulting representations are no longer the vectorization or matricization of states, POVM elements, and CPTP maps. Instead, one should treat the collection of descriptions for all gates (and states, POVMs) as a joint description of the entire device. Quoting from [39], "GST is tomography of a novel entity". But what is this novel entity? Based on our theory, the novel entity is the extrinsic behavior of the world models, i.e., the collection of probability distributions over measurement outcomes under experiments. In particular, GST learns the extrinsic behavior of world models that satisfy the assumption that we can identify a set of (composed) states and a set of (composed) POVM elements, such that both sets span the entire d-dimensional quantum state space. When the assumption is not satisfied, Theorem 3 shows that the complexity of learning extrinsic behaviors can be as bad as running all experiments. Building on this realization, we can rigorously study the performance of different GST protocols: we simply look at whether the protocols are able to accurately predict new experimental outcomes. In Theorem 4, we build on this realization and develop a rigorous algorithm that can provably predict future experiments.

K.2. Quantum state/process/measurement tomography

The settings studied in quantum state, process, and measurement tomography are special cases of our theory. In quantum state tomography, we are learning a world model such that the CPTP maps and POVM measurements are perfect. Quantum state tomography focuses on learning unknown quantum state ρ to high accuracy, usually in trace norm or fidelity. See e.g. [2, 3, 21, 27] and references within. In quantum measurement tomography [12], we are learning a world model such that the state preparations and the the CPTP maps are perfect. The goal of quantum measurement tomography is to learn the descriptions of the POVMs. Finally, in quantum process tomography [38], we are learning a world model such that the state preparations and the POVM measurements are perfect. The purpose of quantum process tomography is to learn the full description of some unknown quantum processes. A subset of quantum processes that have been actively studied recently are the Pauli channels [17, 18, 24], which are often considered to be the major noise sources in quantum computers.

L. Quantum advantage with noisy devices that cannot be fully learned

L.1. Setting

We are given an $(2n)$-qubit device with unknown operations.

L.1.a. State preparation and measurements

We can prepare an unknown product state given by

$$\rho_1 \otimes \ldots \otimes \rho_{2n}. \tag{L1}$$

We can perform an unknown product measurement given by a POVM with 2^{2n} outcomes,

$$\left\{ \bigotimes_{i=1}^{2n} M_{B_i}^{(i)} \right\}_{B \in \{0,1\}^{2n}}, \tag{L2}$$

where $M_{B_i}^{(i)}, \forall i = 1, \ldots, 2n$ are 2×2 positive-semidefinite matrices.

L.1.b. Operations

In the device, we can apply layers of non-overlapping single- and two-qubit gates. We consider the $2n$ qubits to be represented by an $n \times 2$ grid. The grid coordinate of the i-th qubit (i from 1 to n) is $(\lceil i/2 \rceil, i \bmod 2)$. Single-qubit gates can be applied at every grid point. But two-qubit gates can only be applied at edges on the $n \times 2$ grid. Each single-qubit gate is an unknown CPTP map on the corresponding single qubit. And each two-qubit gate is an unknown CPTP map on the corresponding two qubits. The unknown CPTP map implemented by each gate depend on the presence or absence of all the other gates applied at the same layer.

Finally, we assume that we can load some unknown n-qubit quantum state ρ on the n qubits in the left hand side of the $n \times 2$ grid. Suppose σ is the quantum state on the $2n$ qubits. After loading ρ, the state on the $2n$ qubits becomes $\rho \otimes \mathrm{tr}_{\text{left side}}(\sigma)$.

L.2. Algorithm for partially learning the device

In this section, we will partially learn some of the operations in the unknown quantum device. We will obtain some descriptions that are useful for using the quantum device to achieve advantage in performing entangled data analysis. The descriptions do not fully characterize the device. But the lack of full characterization enables us to work under a more general setting.

L.2.a. Experiments and loss functions: Single-qubit

We perform an optimization to find nine single-qubit gates for every qubit, denoted as $g_{i,k}$ for all $i = 1, \ldots, 2n, k = 0, \ldots, 8$, by minimizing a loss function. In particular, after choosing $g_{i,k}$, we conduct the following experiments.

We consider the same single-qubit gates on the left hand side and the right hand side of the $n \times 2$ qubit grid. For each of $k_1, k_1' \in \{0, \ldots, 5\}, k_2, k_2' \in \{6, 7, 8\}$, we prepare the unknown product state $\rho_1 \otimes \ldots \otimes \rho_{2n}$, apply a layer of single-qubit gates, followed by another layer of single-qubit gates, then measure using the unknown product measurement. In the first layer of the single-qubit gates, we apply g_{i,k_1} for qubit i on the left and apply $g_{i,k_1'}$ for qubit i on the right. In the second layer of the single-qubit gates, we apply g_{i,k_2} for qubit i on the left and apply $g_{i,k_2'}$ for qubit i on the right. For

each k_1, k_1', k_2, k_2' and each qubit i, we estimate the probability for obtaining the outcome $b \in \{0,1\}$, denoted as $\hat{p}_{k_1,k_1',k_2,k_2',i,b}$. From a total of $\mathcal{O}(\log(n/\delta)/\eta_0^2)$, the estimate $\hat{p}_{k_1,k_1',k_2,k_2',i,b}$ is equal to the true probability up to η_0 error for all $k_1, k_1', k_2, k_2', i, b$ with a probability at least $1-\delta$. We now define a loss function based on the estimated values,

$$\eta_1 = \max_{k_1,k_1',k_2,k_2',i,b} \left| \hat{p}_{k_1,k_1',k_2,k_2',i,b} - f(k_1, k_1', k_2, k_2', i, b) \right|. \tag{L3}$$

The function $f(k_1, k_1', k_2, k_2', i, b)$ is defined as follows. For qubit i on the left, we have

$$f(k_1, k_1', k_2, k_2', i, b) = \begin{cases} 1/2 & \lfloor k_1/2 \rfloor \neq (k_2 - 6) \\ 1 & \lfloor k_1/2 \rfloor = (k_2 - 6),\ k_1 \equiv b \,(\mathrm{mod} 2) \\ 0 & \lfloor k_1/2 \rfloor = (k_2 - 6),\ k_1 \not\equiv b \,(\mathrm{mod} 2) \end{cases} \tag{L4}$$

For qubit i on the right, we have

$$f(k_1, k_1', k_2, k_2', i, b) = \begin{cases} 1/2 & \lfloor k_1'/2 \rfloor \neq (k_2' - 6) \\ 1 & \lfloor k_1'/2 \rfloor = (k_2' - 6),\ k_1' \equiv b \,(\mathrm{mod} 2) \\ 0 & \lfloor k_1'/2 \rfloor = (k_2' - 6),\ k_1' \not\equiv b \,(\mathrm{mod} 2) \end{cases} \tag{L5}$$

We optimize over the selection of gates such that the loss function η_1 is as small as possible.

L.2.b. Rigorous guarantee: Single-qubit

Given an estimation error η_0 and the loss function η_1, we can approximately learn the following states and measurements. We combine the first layer of single-qubit gates determined by k_1, k_1' and the unknown product state to create a new set of product states, denoted as

$$\rho_1^{(k_1,k_1')} \otimes \ldots \otimes \rho_{2n}^{(k_1,k_1')}, \tag{L6}$$

for all $k_1, k_1' \in \{0,1,2,3,4,5\}$. We also combine the last layer of single-qubit gates determined by k_2, k_2' and the unknown product measurement to create a new set of product measurements,

$$\left\{ \bigotimes_{i=1}^{2n} M_{B_i}^{(i,k_2,k_2')} \right\}_{B \in \{0,1\}^{2n}}, \tag{L7}$$

for all $k_2, k_2' \in \{6,7,8\}$. We have the following characterization.

Lemma L.1 (Single-qubit stabilizer states)**.** *For any $i = 1, \ldots, 2n$, there exists a unitary or anti-unitary transformation U_i on qubit i, such that the following holds for any k_1, k_1', k_2, k_2', b. For qubit i on the left, we have*

$$\left\| \rho_i^{(k_1,k_1')} - U_i \sigma_{k_1} U_i^{-1} \right\|_1 \leq \mathcal{O}(\eta_0 + \eta_1), \tag{L8}$$

$$\left\| M_b^{(i,k_2,k_2')} - U_i \sigma_{2(k_2-6)+b} U_i^{-1} \right\|_1 \leq \mathcal{O}(\eta_0 + \eta_1). \tag{L9}$$

For qubit i on the right, we have

$$\left\| \rho_i^{(k_1,k_1')} - U_i \sigma_{k_1'} U_i^{-1} \right\|_1 \leq \mathcal{O}(\eta_0 + \eta_1), \tag{L10}$$

$$\left\| M_b^{(i,k_2,k_2')} - U_i \sigma_{2(k_2'-6)+b} U_i^{-1} \right\|_1 \leq \mathcal{O}(\eta_0 + \eta_1), \tag{L11}$$

where the pure states σ_x for $x = 0, \ldots, 5$ are given by

$$|0\rangle\langle 0|,\ |1\rangle\langle 1|,\ |+\rangle\langle +|,\ |-\rangle\langle -|,\ |y+\rangle\langle y+|,\ |y-\rangle\langle y-|. \tag{L12}$$

Proof. Without loss of generality, we consider qubit i on the left hand side of the $n \times 2$ grid. The proof is the same for qubits on the right side except that we need to swap $k_1 \leftrightarrow k_1', k_2 \leftrightarrow k_2'$. From the estimate $\hat{p}_{k_1,k_1',k_2,k_2',i,b}$ and the definition of the loss function η_1,

$$\left|\mathrm{tr}\left(M_b^{(i,k_2,k_2')}\rho_i^{(k_1,k_1')}\right) - f(k_1,k_1',k_2,k_2',i,b)\right| \leq \eta_0 + \eta_1, \tag{L13}$$

for all k_1, k_1', k_2, k_2', b. The above defines an approximate geometry that can be used to infer the underlying operations. This technique is used in Appendix D.5 to learn general d-dimensional systems.

For a fixed choice of k_1', k_2', we have six matrices for $\rho_i^{(k_1,k_1')}$ and six matrices for $M_b^{(i,k_2,k_2')}$ corresponding to different k_1, k_2 and b. Now, given $q = 0, 1, 2$. We denote the two matrices $\rho_i^{(k_1,k_1')}$ associated to $k_1 = 2q+0, 2q+1$ as ρ_0, ρ_1, denote the two matrices $M_b^{(i,k_2,k_2')}$ associated to $k_2 = q$ and $b = 0, 1$ as M_0, M_1, and define $\eta = \eta_0 + \eta_1$. We have the following inequalities from Eq. (L13),

$$\mathrm{tr}(M_0\rho_0) \geq 1-\eta,\ \ \mathrm{tr}(M_0\rho_1) \leq \eta,\ \ \mathrm{tr}(M_1\rho_0) \leq \eta,\ \ \mathrm{tr}(M_1\rho_1) \geq 1-\eta. \tag{L14}$$

Using the fact that M_0, M_1, ρ_0, ρ_1 are positive-semidefinite, $M_0 + M_1 = I$, and $\mathrm{tr}(\rho_0) = \mathrm{tr}(\rho_1) = 1$, there exists two orthogonal pure states $|\psi_0\rangle\langle\psi_0|, |\psi_1\rangle\langle\psi_1|$ and a constant $C > 0$, such that

$$\|M_0 - |\psi_0\rangle\langle\psi_0|\|_1 \leq C\eta, \qquad \|M_1 - |\psi_1\rangle\langle\psi_1|\|_1 \leq C\eta, \tag{L15}$$

$$\|\rho_0 - |\psi_0\rangle\langle\psi_0|\|_1 \leq C\eta, \qquad \|\rho_1 - |\psi_1\rangle\langle\psi_1|\|_1 \leq C\eta. \tag{L16}$$

Hence, we know that the six matrices for $\rho_i^{(k_1,k_1')}$ are approximately pure states. Then we can use the approximate geometry given in Eq. (L13) over pairs of distinct q to show that there exists a constant $C' > 0$ and a unitary or anti-unitary transformation $U_{i,k_1',k_2'}$ such that

$$\left\|\rho_i^{(k_1,k_1')} - U_{i,k_1',k_2'}\sigma_{k_1}U_{i,k_1',k_2'}^{-1}\right\|_1 \leq C'\eta, \tag{L17}$$

$$\left\|M_b^{(i,k_2,k_2')} - U_{i,k_1',k_2'}\sigma_{2(k_2-6)+b}U_{i,k_1',k_2'}^{-1}\right\|_1 \leq C'\eta, \tag{L18}$$

where the pure states σ_x for $x = 0, \ldots, 5$ are given by

$$\sigma_x = |0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|, |y+\rangle\langle y+|, |y-\rangle\langle y-|, \tag{L19}$$

respectively.

We see that $U_{i,k_1',k_2'}$ depends on k_1', k_2'. The last step is to show that we can actually choose a single unitary or anti-unitary transformation U_i. Consider $U_i = U_{i,k_1'=0,k_2'=6}$. For all $x = 0, \ldots, 5$, we have

$$\begin{aligned}
&\left\|U_{i,k_1',k_2'}\sigma_x U_{i,k_1',k_2'}^{-1} - U_i\sigma_x U_i^{-1}\right\|_1 && \text{(L20)}\\
&\leq \left\|U_{i,k_1',k_2'}\sigma_x U_{i,k_1',k_2'}^{-1} - U_{i,0,k_2'}\sigma_x U_{i,0,k_2'}^{-1}\right\|_1 + \left\|U_{i,0,k_2'}\sigma_x U_{i,0,k_2'}^{-1} - U_{i,0,6}\sigma_x U_{i,0,6}^{-1}\right\|_1 && \text{(L21)}\\
&\leq \left\|U_{i,k_1',k_2'}\sigma_x U_{i,k_1',k_2'}^{-1} - \rho_i^{(x,k_1')}\right\|_1 + \left\|\rho_i^{(x,k_1')} - U_{i,0,k_2'}\sigma_x U_{i,0,k_2'}^{-1}\right\|_1 && \text{(L22)}\\
&+ \left\|U_{i,0,k_2'}\sigma_x U_{i,0,k_2'}^{-1} - M_{x \bmod 2}^{(i,k_2,\lfloor x/2\rfloor+6)}\right\|_1 + \left\|M_{x \bmod 2}^{(i,k_2,\lfloor x/2\rfloor+6)} - U_{i,0,6}\sigma_x U_{i,0,6}^{-1}\right\|_1 && \text{(L23)}\\
&\leq 2C'\eta + 2C'\eta = 4C'\eta. && \text{(L24)}
\end{aligned}$$

The second-to-last inequality follows from Eq. (L17) and (L18). Together, we have

$$\left\|\rho_i^{(k_1,k_1')} - U_i\sigma_{k_1}U_i^{-1}\right\|_1 \leq \mathcal{O}(\eta_0+\eta_1), \tag{L25}$$

$$\left\|M_b^{(i,k_2,k_2')} - U_i\sigma_{2(k_2-6)+b}U_i^{-1}\right\|_1 \leq \mathcal{O}(\eta_0+\eta_1), \tag{L26}$$

This concludes the proof. □

L.2.c. Experiments and loss functions: Two-qubit

We also perform another layer of optimization to find two two-qubit gates for every pair of corresponding qubits on the left and right, i.e., qubit $i = 2\ell - 1, 2\ell$, where $\ell = 1, \ldots, n$. We denote the two two-qubit gates as $g^{(2)}_{\ell,s}$ and $g^{(2)}_{\ell,e}$ for each $\ell = 0, \ldots, n$. Similar to before, after choosing $g^{(2)}_{\ell,s}$ and $g^{(2)}_{\ell,e}$, we conduct the following set of experiments.

For each of $k_1, k_1' \in \{0, \ldots, 5\}, k_2, k_2' \in \{6, 7, 8\}, x \in \{s, e\}$, we prepare the unknown product state $\rho_1 \otimes \ldots \otimes \rho_{2n}$, apply a layer of single-qubit gates, a layer of two-qubit gates, followed by another layer of single-qubit gates, then measure using the unknown product measurement. In the first layer of the single-qubit gates, we apply g_{i,k_1} for qubit i on the left and apply $g_{i,k_1'}$ for qubit i on the right. In the middle layer of two-qubit gates, we apply $g^{(2)}_{\ell,x}$ on qubit $2\ell - 1$ on the left and qubit 2ℓ on the right. In the other layer of the single-qubit gates, we apply g_{i,k_2} for qubit i on the left and apply $g_{i,k_2'}$ for qubit i on the right. For each k_1, k_1', k_2, k_2' and each pair of qubits $2\ell - 1, 2\ell$, we estimate the probability for obtaining the two bits $b, b' \in \{0, 1\}$ as the outcome, denoted as $\hat{p}^{(2)}_{k_1,k_1',k_2,k_2',x,\ell,b,b'}$. From a total of $\mathcal{O}(\log(n/\delta)/\eta_0^2)$, the estimate $\hat{p}^{(2)}_{k_1,k_1',k_2,k_2',x,\ell,b,b'}$ is equal to the true probability up to η_0 error for all $k_1, k_1', k_2, k_2', x, \ell, b, b'$ with a probability at least $1 - \delta$. We now define another loss function based on the estimated values,

$$\eta_2 = \max_{k_1,k_1',k_2,k_2',x,\ell,b,b'} \left| \hat{p}^{(2)}_{k_1,k_1',k_2,k_2',x,\ell,b,b'} - h(k_1, k_1', k_2, k_2', x, b, b') \right|. \tag{L27}$$

The function $h(k_1, k_1', k_2, k_2', x, b, b')$ is defined as follows.

$$h(k_1, k_1', k_2, k_2', x, b, b') = \operatorname{tr}\left((\rho^{(\text{out})}_{k_2,b} \otimes \rho^{(\text{out})}_{k_2',b'}) U_x (\rho^{(\text{in})}_{k_1} \otimes \rho^{(\text{in})}_{k_1'}) U_x^\dagger \right), \tag{L28}$$

$$\rho^{(\text{in})}_{k_1} = \sigma_{k_1}, \tag{L29}$$

$$\rho^{(\text{out})}_{k_2,b} = \sigma_{2(k_2-6)+b}, \tag{L30}$$

$$\sigma_0, \ldots, \sigma_5 = |0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|, |y+\rangle\langle y+|, |y-\rangle\langle y-|, \tag{L31}$$

$$U_s, U_e = \text{SWAP}, \text{BELL}, \tag{L32}$$

where SWAP and BELL are two-qubit unitaries, SWAP swaps the left and right qubits, BELL $= (H \otimes I)$ CNOT, CNOT is controlled on the same qubit that the Hadamard H acts on. The entangling operation BELL followed by computational basis measurement is one way to perform Bell measurement.

L.2.d. Rigorous guarantee: Two-qubit

Next, using the bound on the estimation error η_0 and the loss function η_2, we can approximately learn the entangling operations. We can write the middle layer of two-qubit gates as

$$\mathcal{E}_{\ell=1,x} \otimes \ldots \otimes \mathcal{E}_{\ell=n,x}, \tag{L33}$$

for $x \in \{s, e\}$, where $\mathcal{E}_{\ell,x}$ acts on two qubits, $2\ell - 1$ and 2ℓ.

Lemma L.2 (SWAP and Bell measurement)**.** *For each qubit $i = 1, \ldots, 2n$, consider the unitary or anti-unitary transformation U_i on qubit i given in Lemma L.1.*

$$\left\| \mathcal{E}_{\ell,s}(\cdot) - (U_{2\ell-1} \otimes U_{2\ell}) \text{SWAP} (U_{2\ell-1} \otimes U_{2\ell})^{-1} (\cdot) (U_{2\ell-1} \otimes U_{2\ell}) \text{SWAP} (U_{2\ell-1} \otimes U_{2\ell})^{-1} \right\|_{1\to 1} \tag{L34}$$

$$\leq \mathcal{O}(\eta_0 + \eta_1 + \eta_2), \tag{L35}$$

$$\left\| \mathcal{E}_{\ell,e}(\cdot) - (U_{2\ell-1} \otimes U_{2\ell}) \text{BELL} (U_{2\ell-1} \otimes U_{2\ell})^{-1} (\cdot) (U_{2\ell-1} \otimes U_{2\ell}) \text{BELL}^\dagger (U_{2\ell-1} \otimes U_{2\ell})^{-1} \right\|_{1\to 1} \tag{L36}$$

$$\leq \mathcal{O}(\eta_0 + \eta_1 + \eta_2), \tag{L37}$$

where $\|\mathcal{E}\|_{1\to 1} = \max_{\|X\|_1 = 1} \|\mathcal{E}(X)\|_1$.

Proof. We prove this lemma for SWAP. The proof for BELL is basically the same. We begin by defining the following notations.

$$\rho^{(in)} = (\sigma_{k_1} \otimes \sigma_{k_1'}), \tag{L38}$$
$$\rho^{(out)} = (\sigma_{2(k_2-6)+b} \otimes \sigma_{2(k_2'-6)+b'}), \tag{L39}$$
$$\rho^{(in,U)} = (U_{2\ell-1}\sigma_{k_1}U_{2\ell-1}^{-1}) \otimes (U_{2\ell}\sigma_{k_1'}U_{2\ell}^{-1}), \tag{L40}$$
$$\rho^{(out,U)} = (U_{2\ell-1}\sigma_{2(k_2-6)+b}U_{2\ell-1}^{-1}) \otimes (U_{2\ell}\sigma_{2(k_2'-6)+b'}U_{2\ell}^{-1}), \tag{L41}$$
$$\sigma_0, \ldots, \sigma_5 = |0\rangle\langle 0|, |1\rangle\langle 1|, |+\rangle\langle +|, |-\rangle\langle -|, |y+\rangle\langle y+|, |y-\rangle\langle y-|. \tag{L42}$$

We can use triangle inequality in Eq. (L27) to show that

$$\left|\mathrm{tr}\left(\left(M_b^{2\ell-1,k_2,k_2'} \otimes M_{b'}^{2\ell,k_2,k_2'}\right)\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right) - \mathrm{tr}\left(\rho^{(out)}\mathrm{SWAP}\left(\rho^{(in)}\right)\mathrm{SWAP}\right)\right| \tag{L43}$$
$$\leq \left|\mathrm{tr}\left(\left(M_b^{2\ell-1,k_2,k_2'} \otimes M_{b'}^{2\ell,k_2,k_2'}\right)\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right) - \hat{p}^{(2)}_{k_1,k_1',k_2,k_2',x,\ell,b,b'}\right| \tag{L44}$$
$$+ \left|\hat{p}^{(2)}_{k_1,k_1',k_2,k_2',x,\ell,b,b'} - \mathrm{tr}\left(\rho^{(out)}\mathrm{SWAP}\left(\rho^{(in)}\right)\mathrm{SWAP}\right)\right| \tag{L45}$$
$$\leq \eta_0 + \eta_2. \tag{L46}$$

Then from triangle inequality and Lemma L.1, we have

$$\left|\mathrm{tr}\left(\left(M_b^{2\ell-1,k_2,k_2'} \otimes M_{b'}^{2\ell,k_2,k_2'}\right)\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right) - \mathrm{tr}\left(\rho^{(out,U)}\mathcal{E}_{\ell,s}\left(\rho^{(in,U)}\right)\right)\right| \tag{L47}$$
$$\leq \left|\mathrm{tr}\left(\left(M_b^{2\ell-1,k_2,k_2'} \otimes M_{b'}^{2\ell,k_2,k_2'}\right)\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right) - \mathrm{tr}\left(\rho^{(out,U)}\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right)\right| \tag{L48}$$
$$+ \left|\mathrm{tr}\left(\rho^{(out,U)}\mathcal{E}_{\ell,s}\left(\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'}\right)\right) - \mathrm{tr}\left(\rho^{(out,U)}\mathcal{E}_{\ell,s}\left(\rho^{(in,U)}\right)\right)\right| \tag{L49}$$
$$\leq \left\|\left(M_b^{2\ell-1,k_2,k_2'} \otimes M_{b'}^{2\ell,k_2,k_2'}\right) - \rho^{(out,U)}\right\|_1 + \left\|\rho_{2\ell-1}^{k_1,k_1'} \otimes \rho_{2\ell}^{k_1,k_1'} - \rho^{(in,U)}\right\|_1 \tag{L50}$$
$$\leq \mathcal{O}(\eta_0 + \eta_1). \tag{L51}$$

Hence, by another triangle inequality, we obtain

$$\left|\mathrm{tr}\left(\rho^{(out,U)}\mathcal{E}_{\ell,s}\left(\rho^{(in,U)}\right)\right) - \mathrm{tr}\left(\rho^{(out)}\mathrm{SWAP}\left(\rho^{(in)}\right)\mathrm{SWAP}\right)\right| \leq \mathcal{O}(\eta_0 + \eta_1 + \eta_2). \tag{L52}$$

Because every $(2\times2)\times(2\times2)$ Hermitian matrix can be written as a linear combination of $\sigma_i \otimes \sigma_j, \forall i, j \in \{0, 1, \ldots, 5\}$, we have the following bound by triangle inequality,

$$\left\|\mathcal{E}_{\ell,s}(\cdot) - (U_{2\ell-1} \otimes U_{2\ell})\mathrm{SWAP}(U_{2\ell-1} \otimes U_{2\ell})^{-1}(\cdot)(U_{2\ell-1} \otimes U_{2\ell})\mathrm{SWAP}(U_{2\ell-1} \otimes U_{2\ell})^{-1}\right\|_{1\to1} \tag{L53}$$
$$\leq \mathcal{O}(\eta_0 + \eta_1 + \eta_2). \tag{L54}$$

This concludes the proof. □

Remark 6. If $\eta_0 + \eta_1 + \eta_2$ is smaller than some constant, then $U_{2\ell-1}$ and $U_{2\ell}$ must be both unitary or both anti-unitary, which can be shown by a proof by contradiction.

L.2.e. Putting everything together

When the quantum device is designed perfectly, the loss functions vanish, i.e., $\eta_1 = \eta_2 = 0$. However, when the quantum device is subject to some unknown noise, the loss functions η_1, η_2 will be small but non-zero. Here, we show that when we perform sufficient number of experiments to estimate the loss functions and we find the loss functions to be small, then we can guarantee that the underlying physical operations satisfy a certain form.

From now on, we assume that $\eta_0+\eta_1+\eta_2$ are small enough such that there exists $\epsilon>0$ and a set of unitary or anti-unitary transformations $U_i, \forall i=1,\ldots,2n$ satisfying the following constraints. For all $\ell=1,\ldots,n$, $k_1,k_1'\in\{0,\ldots,5\}$, and $k_2,k_2'\in\{6,7,8\}$, we have

$$\left\|\rho_{2\ell-1}^{(k_1,k_1')}\otimes\rho_{2\ell}^{(k_1,k_1')}-(U_{2\ell-1}\sigma_{k_1}U_{2\ell}^{-1})\otimes(U_{2\ell}\sigma_{k_1'}U_{2\ell}^{-1})\right\|_1\leq\epsilon, \tag{L55}$$

$$\sum_{b,b'\in\{0,1\}}\left\|M_b^{(2\ell-1,k_2,k_2')}\otimes M_{b'}^{(2\ell,k_2,k_2')}-(U_{2\ell-1}\sigma_{2(k_2-6)+b}U_{2\ell}^{-1})\otimes(U_{2\ell}\sigma_{2(k_2'-6)+b'}U_{2\ell}^{-1})\right\|_1\leq\epsilon, \tag{L56}$$

where the pure states σ_x for $x=0,\ldots,5$ are given by

$$|0\rangle\langle 0|,\ |1\rangle\langle 1|,\ |+\rangle\langle +|,\ |-\rangle\langle -|,\ |y+\rangle\langle y+|,\ |y-\rangle\langle y-|. \tag{L57}$$

And we also have

$$\left\|\mathcal{E}_{\ell,s}-(U_{2\ell-1}\otimes U_{2\ell})\mathrm{SWAP}(U_{2\ell-1}\otimes U_{2\ell})^{-1}(\cdot)(U_{2\ell-1}\otimes U_{2\ell})\mathrm{SWAP}(U_{2\ell-1}\otimes U_{2\ell})^{-1}\right\|_{1\to1}\leq\epsilon, \tag{L58}$$

$$\left\|\mathcal{E}_{\ell,e}-(U_{2\ell-1}\otimes U_{2\ell})\mathrm{BELL}(U_{2\ell-1}\otimes U_{2\ell})^{-1}(\cdot)(U_{2\ell-1}\otimes U_{2\ell})\mathrm{BELL}^\dagger(U_{2\ell-1}\otimes U_{2\ell})^{-1}\right\|_{1\to1}\leq\epsilon, \tag{L59}$$

where SWAP and BELL are two-qubit unitaries, SWAP swaps the left and right qubits, BELL $=(H\otimes I)$ CNOT, and $\|\mathcal{E}\|_{1\to1}=\max_{\|X\|_1=1}\|\mathcal{E}(X)\|_1$.

L.3. Task description and quantum advantage

We focus on the task of predicting many incompatible properties in unknown physical systems studied in [9, 28, 30]. We consider an unknown physical system described by an n-qubit separable state ρ. Recall that a separable state is a classical probability mixture over product states. The goal is to learn to predict properties about ρ.

We compare two experimental settings: conventional experiments and quantum-enhanced experiments. In *conventional experiments*, the physicist could perform any POVM measurement on the unknown state ρ to gather classical data. Based on the measurement data, the physicist could adaptively choose the next measurement on ρ to obtain more data. After many rounds of measurements, the physicist combines all measurement outcomes to form a model about the unknown state ρ. In *quantum-enhanced experiments*, the physicist could load multiple copies of the unknown state ρ to a quantum computing system. The physicist can then use the quantum computer to perform quantum data analysis to learn a model about the unknown state ρ. After learning a model of ρ, we will ask the physicist to predict properties about ρ.

We assume that the conventional experiments are perfect. All the POVM measurement could be chosen arbitrarily and are not subject to any noise. In contrast, we consider quantum-enhanced experiments to base only on the noisy and unknown quantum device that we partially learned in Appendix L.2. Despite the imperfection of quantum-enhanced experiments, we can still demonstrate a large quantum advantage.

Theorem L.3 (Advantage with noisy quantum device; Restatement of Theorem 7)**.** *Given $n>0$. Suppose ϵ is the error on each two-qubit operation in the noisy quantum device. There exists a distribution over unknown n-qubit separable states ρ and properties, such that quantum-enhanced experiments using the noisy quantum device only require $N_{\mathrm{Q}}=\mathcal{O}((1/(1-4\epsilon))^{2n})$ experiments to predict accurately, while noiseless conventional experiments require $N_{\mathrm{C}}=\Omega(2^n)$ experiments. This yields a separation of $N_{\mathrm{C}}=\Omega(N_{\mathrm{Q}}^a)$, where $a=-\log(2)/(2\log(1-4\epsilon))=\mathcal{O}(\epsilon)$.*

L.4. A class of states and properties

For certain states and properties, conventional experiments can be very powerful. For example, classical shadow tomography [29] is a class of conventional experiments based on randomized measurements that can be used to make accurate prediction for many properties. Here, we give an example

where there is a large separation between conventional experiments with perfect measurements and noisy quantum-enhanced experiments. We consider a distribution over the unknown n-qubit separable states ρ and the properties as follows.

With probability $1/(4^n-1)$, we sample a Pauli observable $P \in \{I,X,Y,Z\}^{\otimes n} \setminus \{I^{\otimes n}\}$. Then with probability $1/2$, we consider the unknown state ρ to be the maximally mixed state $I/2^n$. With probability $1/4$, we consider the unknown state ρ to be

$$\left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right)\left(\frac{I+0.9P}{2^n}\right)\left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right). \tag{L60}$$

With probability $1/4$, we consider the unknown state ρ to be

$$\left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right)\left(\frac{I-0.9P}{2^n}\right)\left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right). \tag{L61}$$

The property we would like to predict is the absolute value of the expectation value of

$$O = \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right) P \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right). \tag{L62}$$

We tell the learning algorithm both the Pauli observable P and the above observable O. When $\rho = I/2^n$, the property is equal to zero. However, when ρ is an alternative state that is not $I/2^n$, the property is equal to 0.9. Hence, making accurate prediction in this task is equivalent to distinguishing if ρ is the maximally mixed state $I/2^n$.

L.4.a. Characterization of the probability distribution

Both the states $I/2^n$ and $(I \pm 0.9P)/2^n$ are separable states and can be represented as a classical probability distribution over tensor products of the single-qubit stabilizer states

$$S = \big\{|0\rangle\langle 0|,\, |1\rangle\langle 1|,\, |+\rangle\langle +|,\, |-\rangle\langle -|,\, |y+\rangle\langle y+|,\, |y-\rangle\langle y-|\big\}. \tag{L63}$$

The physical system ρ is a classical probability mixture over tensor products of $U_{2\ell-1}\sigma_\ell U_{2\ell-1}^{-1}$ for $\ell = 1,\ldots,n$, where σ_ℓ is a single-qubit stabilizer state. Hence, we have

$$\rho = \sum_{\sigma_\ell \in S, \forall \ell = 1,\ldots,n} p(\sigma_1,\ldots,\sigma_n)(U_1\sigma_1 U_1^{-1}) \otimes \ldots \otimes (U_{2n-1}\sigma_n U_{2n-1}^{-1}). \tag{L64}$$

There are multiple distinct probability distributions that lead to the same state ρ. Here, we consider the following classical distribution based on the chosen Pauli observable $P \in \{I,X,Y,Z\}^{\otimes n} \setminus \{I^{\otimes n}\}$ for the ease of analysis.

- ρ is the maximally mixed state: For each $\ell = 1,\ldots,n$, we consider the following. If $P_\ell = I, X, Y$, we choose σ_ℓ from the uniform distribution over $|0\rangle\langle 0|, |1\rangle\langle 1|$. If $P_\ell = Z$, we choose σ_ℓ from the uniform distribution over $|+\rangle\langle +|, |-\rangle\langle -|$.

- ρ is a locally rotated $(I \pm 0.9P)/2^n$: The state $(I \pm 0.9P)/2^n$ is equal to the uniform mixture of the maximally mixed state and the state $(I \pm P)/2^n$. With probability 0.1, we prepare ρ the same way as the maximally mixed state given above. With probability 0.9, we prepare ρ as $(I \pm P)/2^n$. For each $\ell = 1,\ldots,n$, we consider the following. If $P_\ell = I$, we choose a σ_ℓ to be a uniform distribution over $|0\rangle\langle 0|, |1\rangle\langle 1|$. If $P_\ell \neq I$, we choose σ_ℓ to be one of the eigenstate of P_ℓ. When P_ℓ is not the last non-identity Pauli operator, we choose σ_ℓ uniformly at random. When P_ℓ is the last non-identity Pauli operator, we choose σ_ℓ deterministically based on the choice of $\sigma_{\ell'}$ for $\ell' < \ell$ where $P_{\ell'}$ is not an identity. The determinisitic choice satisfies a parity constraint given by the state $(I \pm P)/2^n$.

Other choices of the classical probability distribution that give rise to the same state would yield exactly the same result but the analysis could be slightly more complex.

L.5. Upper bound for noisy quantum-enhanced experiments

We provide a sample complexity upper bound for quantum-enhanced experiments using the noisy quantum device that we partially learned. The quantum-enhanced experiment we implement loads in two copies of the physical system ρ and perform an entangled measurement across the two copies.

L.5.a. Detailed procedure

In the noisy quantum-enhanced experiments, we utilize the noisy quantum device and repeat the following for $N_{\mathrm{Q}}/2$ times.

1. Prepare the initial product state followed by a layer of single-qubit gates $g_{i,0}$ for all qubit $i = 1, \ldots, 2n$.
2. Load the physical system ρ into the left hand side of the $n \times 2$ grid.
3. Apply the entangling layer $\mathcal{E}_{\ell=1,s} \otimes \ldots \otimes \mathcal{E}_{\ell=n,s}$, which is approximately equal to applying SWAP gates between left and right side of the $n \times 2$ grid from Eq. (L58).
4. Load the physical system ρ into the left hand side of the $n \times 2$ grid again.
5. Apply the entangling layer $\mathcal{E}_{\ell=1,e} \otimes \ldots \otimes \mathcal{E}_{\ell=n,e}$, which is approximately equal to rotating the pairs of qubits into a Bell basis from Eq. (L59).
6. Apply a layer of single-qubit gates $g_{i,6}$ for all qubit $i = 1, \ldots, 2n$. Then measure using the unknown product measurement.
7. Store the measurement outcome as $b_{t,i}, \forall i = 1, \ldots, 2n$ for the t-th experiment.

After the $N_{\mathrm{Q}}/2$ experiments using N_{Q} copies of the physical system ρ, when we are given a Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$, we compute the following,

$$\Xi \equiv \frac{2}{N_{\mathrm{Q}}} \sum_{t=1}^{N_{\mathrm{Q}}/2} \prod_{\ell=1}^{n} \langle \beta(b_{t,2\ell-1}, b_{t,2\ell}) | P_\ell \otimes P_\ell | \beta(b_{t,2\ell-1}, b_{t,2\ell}) \rangle , \tag{L65}$$

where $|\beta(x, y)\rangle$ is the Bell state,

$$|\beta(x, y)\rangle = \left(\frac{|0, y\rangle + (-1)^x |1, 1 - y\rangle}{\sqrt{2}} \right). \tag{L66}$$

To understand what is happening, consider $U_i = I, \forall i$. When all the operations are perfect, the quantum-enhanced experiment is equivalent to storing two copies of ρ in the $2n$ qubits and measuring every pair of corresponding qubits in the Bell basis. Bell basis simultaneously diagonalizes $P \otimes P$ for all $P \in \{I, X, Y, Z\}^{\otimes n}$. Hence, we can simultaneously predict $\mathrm{tr}((P \otimes P)(\rho \otimes \rho)) = \mathrm{tr}(P\rho)^2$.

L.5.b. Noise analysis

We analyze how noise affects the quantum-enhanced experiments. Suppose the first sample of ρ is

$$(U_1 \sigma_1^A U_1^{-1}) \otimes \ldots \otimes (U_{2n-1} \sigma_n^A U_{2n-1}^{-1}), \tag{L67}$$

and the second sample of ρ is

$$(U_1 \sigma_1^B U_1^{-1}) \otimes \ldots \otimes (U_{2n-1} \sigma_n^B U_{2n-1}^{-1}). \tag{L68}$$

Let us focus on a pair of qubits $2\ell-1, 2\ell$ for $\ell = 1, \ldots, n$. From Eq. (L55), after the first two steps, the pair of qubits is in a state $\rho^{(2\ell-1,2\ell),a}$ with

$$\left\| \rho^{(2\ell-1,2\ell),a} - (U_{2\ell-1}\sigma_\ell^A U_{2\ell-1}^{-1}) \otimes (U_{2\ell}|0\rangle\langle 0|U_{2\ell}^{-1}) \right\|_1 \leq \epsilon. \tag{L69}$$

After the third and fourth step, the pair of qubits is now in a state $\rho^{(2\ell-1,2\ell),b}$ with

$$\left\| \rho^{(2\ell-1,2\ell),b} - (U_{2\ell-1}\sigma_\ell^B U_{2\ell-1}^{-1}) \otimes (U_{2\ell}\sigma_\ell^A U_{2\ell}^{-1}) \right\|_1 \leq 2\epsilon \tag{L70}$$

using Eq. (L58). After the fifth step, the two-qubit state is now $\rho^{(2\ell-1,2\ell),c}$. From Eq. (L59), we have

$$\left\| \rho^{(2\ell-1,2\ell),c} - (U_{2\ell-1} \otimes U_{2\ell})\mathrm{BELL}(\sigma_\ell^B \otimes \sigma_\ell^A)\mathrm{BELL}^\dagger(U_{2\ell-1}^{-1} \otimes U_{2\ell}^{-1}) \right\|_1 \leq 3\epsilon. \tag{L71}$$

In the sixth and seventh step, we measure the two-qubit state $\rho^{(2\ell-1,2\ell),c}$ with a two-qubit product POVM. The two-qubit product POVM is given by

$$\{M_0^{(2\ell-1,6,6)} \otimes M_0^{(2\ell,6,6)}, M_0^{(2\ell-1,6,6)} \otimes M_1^{(2\ell,6,6)}, M_1^{(2\ell-1,6,6)} \otimes M_0^{(2\ell,6,6)}, M_1^{(2\ell-1,6,6)} \otimes M_1^{(2\ell,6,6)}\}, \tag{L72}$$

where the approximation error is given in Eq. (L56).

We now combine with the classical post-processing in Eq. (L65). Given a Pauli observable $P \in \{I, X, Y, Z\}^{\otimes n}$. For each experiment, we can show that we are measuring the two-qubit state $\rho^{(2\ell-1,2\ell),c}$ with an observable

$$O^{(2\ell-1,2\ell)} \equiv \sum_{x,y\in\{0,1\}} \langle\beta(x,y)| P_\ell \otimes P_\ell |\beta(x,y)\rangle M_x^{(2\ell-1,6,6)} \otimes M_y^{(2\ell,6,6)}. \tag{L73}$$

The observable $O^{(2\ell-1,2\ell)}$ differs from the ideal observable

$$\sum_{x,y\in\{0,1\}} \langle\beta(x,y)| P_\ell \otimes P_\ell |\beta(x,y)\rangle (U_{2\ell-1}|x\rangle\langle x|U_{2\ell-1}^{-1}) \otimes (U_{2\ell}|y\rangle\langle y|U_{2\ell}^{-1}) \tag{L74}$$

by at most ϵ error in $\|\cdot\|_\infty$. Using triangle inequality, we have

$$\left| \mathrm{tr}\left(O^{(2\ell-1,2\ell)}\rho^{(2\ell-1,2\ell),c}\right) - \sum_{x,y} \langle\beta(x,y)| P_\ell \otimes P_\ell |\beta(x,y)\rangle \langle\beta(x,y)| (\sigma_\ell^B \otimes \sigma_\ell^A) |\beta(x,y)\rangle \right| \leq 4\epsilon. \tag{L75}$$

Because Bell basis simultaneously diagonalizes $\{I \otimes I, X \otimes X, Y \otimes Y, Z \otimes Z\}$, we have

$$\sum_{x,y} \langle\beta(x,y)| P_\ell \otimes P_\ell |\beta(x,y)\rangle \langle\beta(x,y)| (\sigma_\ell^B \otimes \sigma_\ell^A) |\beta(x,y)\rangle = \mathrm{tr}(P_\ell\sigma_\ell^B)\,\mathrm{tr}(P_\ell\sigma_\ell^A). \tag{L76}$$

Recall that there exists a CPTP map $\mathcal{E}_\ell$ such that $\rho^{(2\ell-1,2\ell),c} = \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B)$. Hence, we have

$$\left| \mathrm{tr}\left(O^{(2\ell-1,2\ell)}\mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B)\right) - \mathrm{tr}(P_\ell\sigma_\ell^A)\,\mathrm{tr}(P_\ell\sigma_\ell^B) \right| \leq 4\epsilon. \tag{L77}$$

Furthermore, when $P_\ell = I$, we have

$$\mathrm{tr}\left(O^{(2\ell-1,2\ell)}\mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B)\right) = \mathrm{tr}(P_\ell\sigma_\ell^A)\,\mathrm{tr}(P_\ell\sigma_\ell^B) = 1. \tag{L78}$$

The two characterizations above will be crucial for the sample complexity analysis.

L.5.c. Sample complexity analysis

We use the probability distribution given in Appendix L.4.a to analyze the expectation value of Ξ in Eq. (L65). Combining with Eq. (L64), we have

$$\mathbb{E}[\Xi] = \sum_{\substack{\sigma_\ell^A \in S, \\ \forall \ell = 1,\ldots,n}} \sum_{\substack{\sigma_\ell^B \in S, \\ \forall \ell = 1,\ldots,n}} p(\sigma_1^A, \ldots, \sigma_n^A) p(\sigma_1^B, \ldots, \sigma_n^B) \prod_{\ell=1}^{n} \operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right). \tag{L79}$$

We separate the analysis into two cases.

- ρ is the maximally mixed state: We only consider σ^A, σ^B that appear with nonzero probability. For all $\ell = 1, \ldots, n$, we consider whether P_ℓ is equal to I. If $P_\ell \neq I$, then we have $\operatorname{tr}(P_\ell \sigma_\ell^A) \operatorname{tr}(P_\ell \sigma_\ell^B) = 0$, hence

$$|\operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right)| \leq 4\epsilon \tag{L80}$$

  from Eq. (L77). If $P_\ell = I$, from Eq. (L78), we have

$$\operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) = 1. \tag{L81}$$

  Together, we have

$$\left| \prod_{\ell=1}^{n} \operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) \right| \leq (4\epsilon)^{\#(P_\ell \neq I)}, \tag{L82}$$

  where $\#(P_\ell \neq I)$ is the number of $P_\ell, \forall \ell = 1, \ldots, n$ that is not equal to identity I. Therefore we conclude that

$$|\mathbb{E}[\Xi]| \leq (4\epsilon)^{\#(P_\ell \neq I)}. \tag{L83}$$

- ρ is a locally rotated $(I \pm 0.9P)/2^n$: There is a probability of 0.19 such that at least one of σ^A, σ^B is sampled according to the distribution for the maximally mixed state. The same analysis as the case for maximally mixed state shows that we have

$$\left| \prod_{\ell=1}^{n} \operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) \right| \leq (4\epsilon)^{\#(P_\ell \neq I)}. \tag{L84}$$

  For a probability of 0.81, we have both σ^A and σ^B are sampled according to the probability distribution for $(I \pm P)/2^n$ defined in Appendix L.4.a. We focus on σ^A and σ^B that occur with non-zero probability. We consider all $\ell = 1, \ldots, n$. We again separate into two cases: $P_\ell \neq I$ and $P_\ell = I$.

  - If $P_\ell \neq I$, then we have $\operatorname{tr}(P_\ell \sigma_\ell^A) \operatorname{tr}(P_\ell \sigma_\ell^B) \in \{1, -1\}$. If $\operatorname{tr}(P_\ell \sigma_\ell^A) \operatorname{tr}(P_\ell \sigma_\ell^B) = 1$, then

$$\operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) \geq 1 - 4\epsilon \tag{L85}$$

    from Eq. (L77). In contrast, if $\operatorname{tr}(P_\ell \sigma_\ell^A) \operatorname{tr}(P_\ell \sigma_\ell^B) = -1$, then

$$\operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) \leq -1 + 4\epsilon. \tag{L86}$$

  - If $P_\ell = I$, from Eq. (L78), we have

$$\operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B) \right) = 1. \tag{L87}$$

Furthermore, the parity constraint in the probability distribution over σ^A, σ^B given in Appendix L.4.a shows that

$$\prod_{\ell=1}^{n} \operatorname{tr}(P_\ell \sigma_\ell^A) \operatorname{tr}(P_\ell \sigma_\ell^B) = 1. \tag{L88}$$

Hence, we have

$$\prod_{\ell=1}^{n} \operatorname{tr}\left(O^{(2\ell-1,2\ell)} \mathcal{E}_\ell(\sigma_\ell^A \otimes \sigma_\ell^B)\right) \geq (1-4\epsilon)^{\#(P_\ell \neq I)}. \tag{L89}$$

Combining with Eq. (L84), we can conclude that

$$\mathbb{E}[\Xi] \geq 0.81(1-4\epsilon)^{\#(P_\ell \neq I)} - 0.19(4\epsilon)^{\#(P_\ell \neq I)}. \tag{L90}$$

Recall from Eq. (L65), Ξ is the average over $N_{\mathrm{Q}}/2$ independent random variables bounded between $[-1, 1]$. Hence, by Hoeffding's inequality, we need

$$N_{\mathrm{Q}} = \mathcal{O}(\log(1/\delta)/\tilde{\epsilon}^2) \tag{L91}$$

to estimate $\mathbb{E}[\Xi]$ to $\tilde{\epsilon}$ error with probability at least $1-\delta$. In order to distinguish between ρ being the maximally mixed state or not, we need to estimate $\mathbb{E}[\Xi]$ to an error of at most

$$0.81(1-4\epsilon)^{\#(P_\ell \neq I)} - 0.19(4\epsilon)^{\#(P_\ell \neq I)} - (4\epsilon)^{\#(P_\ell \neq I)}. \tag{L92}$$

For ϵ less than a constant and n sufficiently large, the above function is minimized at $\#(P_\ell \neq I) = n$. In order to predict accurately with a probability at least 0.99, the sample complexity for the noisy quantum-enhanced experiment is

$$N_{\mathrm{Q}} = \mathcal{O}\left(\frac{1}{(1-4\epsilon)^{2n}}\right). \tag{L93}$$

This concludes the proof for the first part of Theorem L.3.

L.6. Lower bound for noiseless conventional experiments

We give a sample complexity lower bound for conventional experiments based on adaptive POVM measurements. We do not assume the presence of any noise in conventional experiments. The proof uses techniques proposed in [9, 28, 30]. In particular, the proof is closely related to one of the proofs in [28] up to minor changes. We present a concise proof here for completeness.

A learning algorithm using noiseless conventional experiments is a rooted tree. At every node, we perform a POVM on ρ. Based on the POVM outcome, we move to a child node of the node. Because a rank-1 POVM $\{w_b|\psi_b\rangle\langle\psi_b|\}_b$ is always at least as powerful as general POVM [9, 30], we will only consider rank-1 POVMs. After $N_{\mathcal{C}}$ experiments, we arrive at a leaf node of the tree at depth $N_{\mathcal{C}}$. Depending on the unknown physical state ρ, the probability distribution over the leaf nodes will be different. We write the probability distribution as

$$p_\rho(\ell) = \prod_{t=1}^{N_{\mathrm{C}}} w_t \langle\psi_t| \rho |\psi_t\rangle, \forall \ell : \text{leaf nodes}, \tag{L94}$$

where $w_1|\psi_1\rangle\langle\psi_1|, \ldots, w_{N_{\mathrm{C}}}|\psi_{\mathrm{C}}\rangle\langle\psi_{\mathrm{C}}|$ are the POVM elements associated to the outcomes of the N_{C} measurements that ends up at the leaf node ℓ.

After the experiments, we obtain a Pauli observable $P \in \{I,X,Y,Z\}^{\otimes n} \setminus \{I^{\otimes n}\}$ and the assocaited observable $O = \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right) P \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right)$. Suppose we can use the conventional experiments to classify between maximally mixed state

$$\rho_{\mathrm{mm}} = I/2^n \tag{L95}$$

and the alternative states

$$\rho_{S,P} = \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right) \left(\frac{I + S0.9P}{2^n}\right) \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right) \tag{L96}$$

under the knowledge of P and $U_i, \forall i$, where $S \in \{\pm 1\}, P \in \{I,X,Y,Z\}^{\otimes n} \setminus \{I^{\otimes n}\}$. Then the average total variation distance between the leaf node distribution of the maximally mixed state and the alternative states $\rho_{s,P}$ must be greater than a constant,

$$\mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\frac{1}{2}\sum_{\ell:\mathrm{leaf}} \left| p_\ell(\rho_{\mathrm{mm}}) - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} p_\ell\left(\rho_{S,P}\right)\right|\right] = \Omega(1). \tag{L97}$$

The expectation over $S \in \{\pm 1\}$ is in the inside because the knowledge of S is not revealed. On the other hand, the expectation over P is on the outside because the knowledge of P is revealed.

We now upper bound the average total variation distance with $N_{\mathcal{C}}$.

$$\mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\frac{1}{2}\sum_{\ell:\mathrm{leaf}} \left| p_\ell(\rho_{\mathrm{mm}}) - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} p_\ell\left(\rho_{S,P}\right)\right|\right] \tag{L98}$$

$$= \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\sum_{\ell:\mathrm{leaf}} \max\left(0,\ p_\ell(\rho_{\mathrm{mm}}) - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} p_\ell\left(\rho_{S,P}\right)\right)\right] \tag{L99}$$

$$= \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\sum_{\ell:\mathrm{leaf}} p_\ell(\rho_{\mathrm{mm}}) \max\left(0,\ 1 - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} \frac{p_\ell\left(\rho_{S,P}\right)}{p_\ell(\rho_{\mathrm{mm}})}\right)\right]. \tag{L100}$$

We lower bound the following term,

$$\mathop{\mathbb{E}}_{S\in\{\pm 1\}} \frac{p_\ell\left(\rho_{S,P}\right)}{p_\ell(\rho_{\mathrm{mm}})} = \mathop{\mathbb{E}}_{S\in\{\pm 1\}} \prod_{t=1}^{N_{\mathrm{C}}} \langle\psi_t| \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}\right) (I + S0.9P) \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right) |\psi_t\rangle \tag{L101}$$

$$= \mathop{\mathbb{E}}_{S\in\{\pm 1\}} \prod_{t=1}^{N_{\mathrm{C}}} \left(1 + 0.9S \langle\tilde{\psi}_t| P |\tilde{\psi}_t\rangle\right) \tag{L102}$$

$$\geq \exp\left[\sum_{t=1}^{N_{\mathrm{C}}} \mathop{\mathbb{E}}_{S\in\{\pm 1\}} \log\left(1 + 0.9S \langle\tilde{\psi}_t| P |\tilde{\psi}_t\rangle\right)\right] \tag{L103}$$

$$= \exp\left[\frac{1}{2}\sum_{t=1}^{N_{\mathrm{C}}} \log\left(1 - 0.81 \langle\tilde{\psi}_t| P |\tilde{\psi}_t\rangle^2\right)\right] \tag{L104}$$

$$= \prod_{t=1}^{N_{\mathrm{C}}} \sqrt{1 - 0.81 \langle\tilde{\psi}_t| P |\tilde{\psi}_t\rangle^2}. \tag{L105}$$

The second line is a definition of $|\tilde{\psi}_t\rangle = \left(\bigotimes_{\ell=1}^{n} U_{2\ell-1}^{-1}\right) |\psi_t\rangle$. The third line uses Jensen's inequality. Hence, we can upper bound the average total variation distance as

$$\mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\frac{1}{2}\sum_{\ell:\mathrm{leaf}} \left| p_\ell(\rho_{\mathrm{mm}}) - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} p_\ell\left(\rho_{S,P}\right)\right|\right] \tag{L106}$$

$$\leq \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\sum_{\ell:\text{leaf}} p_\ell(\rho_{\text{mm}}) \max\left(0,\ 1-\prod_{t=1}^{N_{\text{C}}} \sqrt{1-0.81\left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2}\right)\right] \tag{L107}$$

$$= \sum_{\ell:\text{leaf}} p_\ell(\rho_{\text{mm}}) \left(1- \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \prod_{t=1}^{N_{\text{C}}} \sqrt{1-0.81\left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2}\right). \tag{L108}$$

We can remove the $\max(0,\cdot)$ because $1-\prod_{t=1}^{N_{\text{C}}} \sqrt{1-0.81\left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2} \geq 0$. We bound the term,

$$\mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \prod_{t=1}^{N_{\text{C}}} \sqrt{1-0.81\left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2} \tag{L109}$$

$$\geq \exp\left[\frac{1}{2}\sum_{t=1}^{N_{\text{C}}} \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \log\left(1-0.81\left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2\right)\right] \tag{L110}$$

$$\geq \exp\left[-1.215\sum_{t=1}^{N_{\text{C}}} \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2\right] \tag{L111}$$

$$\geq 1-1.215\sum_{t=1}^{N_{\text{C}}} \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2 \tag{L112}$$

$$= 1-1.215\sum_{t=1}^{N_{\text{C}}} \frac{1}{2^n+1} = 1-\frac{1.215N_{\text{C}}}{2^n+1}. \tag{L113}$$

The second line uses Jensen's inequality. The third line uses $\log(1-x)\geq -3x, \forall x\in[0,0.94]$. The fourth line uses $\exp(x)\geq 1+x, \forall x\in\mathbb{R}$. The last line uses

$$\mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} P\otimes P = \frac{2^n\text{SWAP}-I\otimes I}{4^n-1}, \tag{L114}$$

hence $\mathbb{E}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left\langle\tilde{\psi}_t\right| P \left|\tilde{\psi}_t\right\rangle^2 = \frac{2^n-1}{4^n-1} = \frac{1}{2^n+1}$. Combining with Eq. (L97), we have

$$\Omega(1) = \mathop{\mathbb{E}}_{P\in\{I,X,Y,Z\}^{\otimes n}\setminus\{I^{\otimes n}\}} \left[\frac{1}{2}\sum_{\ell:\text{leaf}} \left| p_\ell(\rho_{\text{mm}}) - \mathop{\mathbb{E}}_{S\in\{\pm 1\}} p_\ell\left(\rho_{S,P}\right)\right|\right] \tag{L115}$$

$$\leq \sum_{\ell:\text{leaf}} p_\ell(\rho_{\text{mm}}) \frac{1.215N_{\text{C}}}{2^n+1} = \frac{1.215N_{\text{C}}}{2^n+1}. \tag{L116}$$

Thus we arrive at the desired lower bound,

$$N_{\text{C}} = \Omega(2^n), \tag{L117}$$

which concludes the proof of Theorem L.3.

[1] D. Aharonov, J. Cotler, and X.-L. Qi. Quantum algorithmic measurement. *Nature Communications*, 13(1):1–9, 2022.

[2] K. Banaszek, M. Cramer, and D. Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.

[3] R. Blume-Kohout. Optimal, reliable estimation of quantum states. *New Journal of Physics*, 12(4):043034, 2010.

[4] R. Blume-Kohout, J. K. Gamble, E. Nielsen, K. Rudinger, J. Mizrahi, K. Fortier, and P. Maunz. Demonstration of qubit operations below a rigorous fault tolerance threshold with gate set tomography. *Nature communications*, 8(1):1–13, 2017.

[5] J. Bourgain and A. Gamburd. A spectral gap theorem in SU(d). *Journal of the European Mathematical Society*, 14(5):1455–1511, 2012.

[6] J. C. Bridgeman and C. T. Chubb. Hand-waving and interpretive dance: an introductory course on tensor networks. *Journal of physics A: Mathematical and theoretical*, 50(22):223001, 2017.

[7] R. Brieger, I. Roth, and M. Kliesch. Compressive gate set tomography. *arXiv preprint arXiv:2112.05176*, 2021.

[8] C. Brif, R. Chakrabarti, and H. Rabitz. Control of quantum phenomena: past, present and future. *New Journal of Physics*, 12(7):075008, 2010.

[9] S. Chen, J. Cotler, H.-Y. Huang, and J. Li. Exponential separations between learning with and without quantum memory. *To appear in FOCS*, 2021.

[10] S. Chen, S. Zhou, A. Seif, and L. Jiang. Quantum advantages for Pauli channel estimation. *arXiv preprint arXiv:2108.08488*, 2021.

[11] C. Dankert, R. Cleve, J. Emerson, and E. Livine. Exact and approximate unitary 2-designs and their application to fidelity estimation. *Physical Review A*, 80(1), Jul 2009.

[12] G. M. D'Ariano, M. G. Paris, and M. F. Sacchi. Quantum tomography. *Advances in Imaging and Electron Physics*, 128:206–309, 2003.

[13] C. M. Dawson and M. A. Nielsen. The Solovay-Kitaev algorithm. *arXiv preprint quant-ph/0505030*, 2005.

[14] S. Endo, S. C. Benjamin, and Y. Li. Practical quantum error mitigation for near-future applications. *Physical Review X*, 8(3):031027, 2018.

[15] Erik, L. Saldyt, Rob, tjproct, J. Gross, sserita, kmrudin, T. L. Scholten, colibri coruscans, kevincyoung, msarovar, coreyostrove, jordanh6, D. Nadlinger, L. N. Maurer, pyIonControl, and R. Blume-Kohout. pygstio/pygsti: Version 0.9.10. Zenodo, Oct. 2021.

[16] S. T. Flammia. Averaged circuit eigenvalue sampling. *arXiv preprint arXiv:2108.05803*, 2021.

[17] S. T. Flammia and R. O'Donnell. Pauli error estimation via population recovery. *arXiv:2105.02885*, 2021.

[18] S. T. Flammia and J. J. Wallman. Efficient estimation of Pauli channels. *ACM Transactions on Quantum Computing*, 1(1):1–32, 2020.

[19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

[20] D. Greenbaum. Introduction to quantum gate set tomography. *arXiv preprint arXiv:1509.02921*, 2015.

[21] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[22] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, 2020.

[23] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Inf. Theory*, 63(9):5628–5641, 2017.

[24] R. Harper, S. T. Flammia, and J. J. Wallman. Efficient learning of quantum noise. *Nature Physics*, 16(12):1184–1188, 2020.

[25] A. W. Harrow, B. Recht, and I. L. Chuang. Efficient discrete approximations of quantum gates. *Journal of Mathematical Physics*, 43(9):4445–4451, 2002.

[26] J. Helsen, M. Ioannou, I. Roth, J. Kitzinger, E. Onorati, A. H. Werner, and J. Eisert. Estimating gate-set properties from random sequences. *arXiv preprint arXiv:2110.13178*, 2021.

[27] Z. Hradil. Quantum-state estimation. *Physical Review A*, 55(3):R1561, 1997.

[28] H.-Y. Huang, M. Broughton, J. Cotler, S. Chen, J. Li, M. Mohseni, H. Neven, R. Babbush, R. Kueng, J. Preskill, et al. Quantum advantage in learning from experiments. *arXiv preprint arXiv:2112.00778*, 2021.

[29] H.-Y. Huang, R. Kueng, and J. Preskill. Predicting many properties of a quantum system from very few measurements. *Nat. Phys.*, 16:1050—1057, 2020.

[30] H.-Y. Huang, R. Kueng, and J. Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.*, 126:190505, 2021.

[31] A. Kandala, K. Temme, A. D. Córcoles, A. Mezzacapo, J. M. Chow, and J. M. Gambetta. Error mitigation extends the computational reach of a noisy quantum processor. *Nature*, 567(7749):491–495, 2019.

[32] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[33] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Ohki. Robust extraction of tomographic information via randomized benchmarking. *Physical Review X*, 4(1):011050, 2014.

[34] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland. Randomized benchmarking of quantum gates. *Physical Review A*, 77(1):012307, 2008.

[35] E. Magesan, J. M. Gambetta, and J. Emerson. Scalable and robust randomized benchmarking of quantum processes. *Physical review letters*, 106(18):180504, 2011.

[36] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. Da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, et al. Efficient measurement of quantum gate error by interleaved

randomized benchmarking. *Physical review letters*, 109(8):080505, 2012.
[37] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. The MIT Press, 2018.
[38] M. Mohseni, A. T. Rezakhani, and D. A. Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.
[39] E. Nielsen, J. K. Gamble, K. Rudinger, T. Scholten, K. Young, and R. Blume-Kohout. Gate set tomography. *arXiv preprint arXiv:2009.07301*, 2020.
[40] T. Proctor, K. Rudinger, K. Young, M. Sarovar, and R. Blume-Kohout. What randomized benchmarking actually measures. *Physical review letters*, 119(13):130502, 2017.
[41] I. Roth, R. Kueng, S. Kimmel, Y.-K. Liu, D. Gross, J. Eisert, and M. Kliesch. Recovering quantum gates from few average gate fidelities. *Physical review letters*, 121(17):170502, 2018.
[42] M. F. Sacchi. Optimal discrimination of quantum operations. *Physical Review A*, 71(6):062340, 2005.
[43] P. Selinger. Efficient Clifford+T approximation of single-qubit operators. *arXiv preprint arXiv:1212.6253*, 2012.
[44] A. Strikis, D. Qin, Y. Chen, S. C. Benjamin, and Y. Li. Learning-based quantum error mitigation. *PRX Quantum*, 2(4):040330, 2021.
[45] T. Surawy-Stepney, J. Kahn, R. Kueng, and M. Guta. Projected least-squares quantum process tomography. *arXiv preprint arXiv:2107.01060*, 2021.
[46] K. Temme, S. Bravyi, and J. M. Gambetta. Error mitigation for short-depth quantum circuits. *Physical review letters*, 119(18):180509, 2017.
[47] P. Varjú. Random walks in compact groups. *arXiv: Group Theory*, 2012.
[48] J. J. Wallman and J. Emerson. Noise tailoring for scalable quantum computation via randomized compiling. *Physical Review A*, 94(5):052325, 2016.
[49] Z. Webb. The clifford group forms a unitary 3-design. *arXiv preprint arXiv:1510.02769*, 2015.
[50] H. M. Wiseman and G. J. Milburn. *Quantum measurement and control*. Cambridge university press, 2009.
[51] T. H. Yang, T. Vértesi, J.-D. Bancal, V. Scarani, and M. Navascués. Robust and versatile black-box certification of quantum devices. *Physical review letters*, 113(4):040401, 2014.
[52] H. Zhu, R. Kueng, M. Grassl, and D. Gross. The clifford group fails gracefully to be a unitary 4-design. *arXiv preprint arXiv:1609.08172*, 2016.