

Exponential separations between learning with and without quantum memory

Sitan Chen*
UC Berkeley

Jordan Cotler†
Harvard University

Hsin-Yuan Huang‡
Caltech

Jerry Li§
Microsoft Research

November 19, 2021

Abstract

We study the power of quantum memory for learning properties of quantum systems and dynamics, which is of great importance in physics and chemistry. Many state-of-the-art learning algorithms require access to an additional external quantum memory. While such a quantum memory is not required *a priori*, in many cases, algorithms that do not utilize quantum memory require much more data than those which do. We show that this trade-off is inherent in a wide range of learning problems. Our results include the following:

- We show that to perform shadow tomography on an n -qubit state ρ with M observables, any algorithm without quantum memory requires $\tilde{\Omega}(\min(M, 2^n))$ samples of ρ in the worst case. Up to logarithmic factors, this matches the upper bound of [HKP20], and completely resolves an open question in [Aar18, AR19].
- We establish exponential separations between algorithms with and without quantum memory for purity testing, distinguishing scrambling and depolarizing evolutions, as well as uncovering symmetry in physical dynamics. Our separations improve and generalize prior work of [ACQ21] by allowing for a broader class of algorithms without quantum memory.
- We give the first tradeoff between quantum memory and sample complexity. More precisely, we prove that to estimate absolute values of all n -qubit Pauli observables, algorithms with $k < n$ qubits of quantum memory require at least $\Omega(2^{(n-k)/3})$ samples, but there is an algorithm using n -qubit quantum memory which only requires $\mathcal{O}(n)$ samples.

The separations we show are sufficiently large and could already be evident, for instance, with tens of qubits. This provides a concrete path towards demonstrating real-world advantage for learning algorithms with quantum memory.

*Email: sitanc@berkeley.edu This work was supported in part by NSF Award 2103300, NSF CAREER Award CCF-1453261, NSF Large CCF-1565235 and Ankur Moitra's ONR Young Investigator Award. Part of this work was completed while visiting the Simons Institute for the Theory of Computing.

†Email: jcotler@fas.harvard.edu. This work is supported by a Junior Fellowship from the Harvard Society of Fellows, as well as in part by the Department of Energy under grant DE-SC0007870.

‡Email: hsinyuan@caltech.edu. This work is supported by the J. Yang & Family Foundation and the Google PhD Fellowship.

§Email: jerrli@microsoft.com. Part of this work was completed while visiting the Simons Institute for the Theory of Computing.

Contents

1	Introduction	1
1.1	Our Results	2
1.1.1	Learning physical systems	2
1.1.2	Smooth tradeoffs for learning with quantum memory	4
1.1.3	Learning quantum dynamics	5
2	Technical Overview	6
2.1	Tree representation	7
2.2	Edge-based analysis	8
2.3	Path-based analysis	9
3	Related Work	12
4	Preliminaries	13
4.1	Tail bounds	13
4.2	Basic results in quantum information theory	13
4.3	Primer on information-theoretic lower bounds	15
4.4	Learning with and without quantum memory	16
4.5	Review of tensor network diagrams	18
5	Exponential separations in learning quantum states	25
5.1	Shadow tomography	27
5.1.1	Lower bound for random observables	28
5.1.2	Lower bound for Pauli observables	31
5.1.3	Upper bound	31
5.2	Testing purity of the quantum state	32
5.2.1	Lower bound	32
5.2.2	Upper bound	34
6	Exponential separation with bounded quantum memory	34
6.1	Lower bound with k -qubit of quantum memory	35
6.2	Upper Bound with n -qubit of quantum memory	39
7	Exponential separations in learning quantum channels	39
7.1	Prerequisites	39
7.2	Review of the Weingarten calculus	41
7.2.1	Haar averaging over $U(d)$	42
7.2.2	Haar averaging over $O(d)$	43
7.2.3	Haar averaging over $\text{Sp}(d/2)$	45
7.3	Depolarizing channel versus random unitary	46
7.3.1	Lower bound without quantum memory	46
7.3.2	Proof of Theorem 7.9	47
7.3.3	Proof of Theorems 7.10 and 7.11	54
7.3.4	Corollaries involving state distinction	66
7.3.5	Upper bound without quantum memory	66
7.3.6	Upper bound with quantum memory	67
7.4	Symmetry distinction problem	67

7.4.1	Lower bound	67
7.4.2	Upper bound	68

1 Introduction

Learning properties of the physical world is at the heart of many scientific disciplines, including physics, chemistry, and material science. The suite of emerging quantum technology, from quantum sensors, quantum computers, to quantum memories for retrieving, processing, and storing quantum information, provides the potential to significantly improve upon what we could achieve with existing technology. Because the physical world is quantum-mechanical, it is natural to expect quantum information technology to provide an advantage in various tasks. Yet, the potential remains to be justified both experimentally and theoretically.

In this work, we establish provable advantages in a wide range of tasks for predicting properties of physical systems and dynamics. We consider *learning algorithms with external quantum memory*, namely those that can retrieve quantum data from each experiment, store the data in quantum memory, and compute on the quantum memory in an entangled fashion (see Definitions 4.16, 4.18, and 4.20). This class of algorithms provides an abstract model for the potential experiments that a scientist assisted by quantum information technology could perform in the future.

We also consider *learning algorithms that do not use external quantum memory*, namely those that can only retrieve classical data by performing measurements after each experiment, store the classical information in classical memory, and process the classical measurement data to learn the properties (see Definitions 4.15 and 4.17). This class of algorithms is an abstract model that encompasses conventional experiments that one could achieve with existing experimental platforms.

Tasks	LB (no q. memory)	UB (no q. memory)	UB (w/ q. memory)
Shadow tomogr.	$\tilde{\Omega}(\min(M, 2^n))$ [Cor 5.7]	$\tilde{\mathcal{O}}(\min(M, 2^n))$ [HKP20]	$\mathcal{O}(n \log(M)^2)$ [BO20]
Shadow tomogr. (P)	$\Omega(2^n)$ [Cor 5.9]	$\mathcal{O}(n2^n)$ [HKP20]	$\mathcal{O}(n)$ [HKP21b]
Purity testing	$\Omega(2^{n/2})$ [Thm 5.11]	$\mathcal{O}(2^{n/2})$ [Thm 5.13]	$\mathcal{O}(1)$ [MdW13]
Depolar. vs unitary	$\Omega(2^{n/3})$ [Thm 7.9]	$\mathcal{O}(2^{n/2})$ [Cor 7.25]	$\mathcal{O}(1)$ [ACQ21]
Classify symmetry	$\Omega(2^{n/3.5})$ [Thm 7.27]	$\mathcal{O}(2^n)$ [Thm 7.28]	$\mathcal{O}(1)$ [ACQ21]
Shadow tomogr. (P) w/ k -qubit memory	Lower bound: $\Omega(2^{(n-k)/3})$ See [Thm 1.4]		

Table 1: A table of the upper and lower bounds on sample/query complexity for learning algorithms with quantum memory (“w/ q. memory”) and without quantum memory (“no q. memory”). For shadow tomography (“Shadow tomogr.”), we consider the sample complexity T needed to achieve a small constant error for estimating expectation values of M observables. The $\tilde{\Omega}$ and $\tilde{\mathcal{O}}$ neglects contributions from $\log(M)$ in the upper and lower bound without quantum memory. Similarly, we consider the sample complexity T for predicting all 4^n Pauli observables to a constant error, denoted by the task: “Shadow tomogr. (P)”. In “Shadow tomogr. (P) w/ k -qubit memory”, we consider a simplified task of predicting absolute value for the expectation of Pauli observables up to a constant error.

In this work, we provide a set of mathematical tools for proving exponential separations between these two classes of learning algorithms. Recently, two independent works [HKP21b, ACQ21] have established exponential separations for some quantum learning tasks. However, their techniques are rather cumbersome and work in more restrictive settings. In this work, we design new, flexible frameworks for proving lower bounds against learning algorithms without quantum memory that strengthen and generalize those presented in [HKP21b, ACQ21]. By using these new ideas, we are able to both significantly tighten their bounds, as well as derive novel bounds for more general and realistic settings that previous techniques could not handle.

Towards quantum advantage with quantum memory. These quantitative and qualitative improvements to the prior known separations have an important consequence for the near-term

goal of demonstrating a provable quantum advantage. As we will explain in more detail in the following, our bounds demonstrate that there exist quantum learning problems on n -qubit states for which there are copy-efficient and gate-efficient quantum algorithms that only require $\mathcal{O}(n)$ bits of quantum memory, but at the same time, any algorithm without quantum memory unconditionally requires $\Omega(2^n)$ copies.

The lower bounds derived in this work are sufficiently tight that this gap is already noticeable at the scale of tens of qubits. In addition, our results guarantee that any protocol with quantum memory of insufficient size must require an exponentially increasing sample complexity; see the last row in Table 1. Together, these results imply that quantum computers with less than a hundred qubits could already assist experimentalists in extracting properties of physical systems and dynamics beyond conventional experiments that do not utilize quantum memory. We believe the implementation of such protocols is an important open experimental question that may be within reach in the coming years. To achieve that, we would need a better understanding on how the presence of noise in the quantum computers affect the advantage studied in this work.

1.1 Our Results

We now informally describe the problems we will consider throughout this paper, as well as our results; we will include pointers to the appropriate sections for the formal theorem statements.

1.1.1 Learning physical systems

We first consider the setting in which there is a physical source that generates copies of an unknown n -qubit state ρ one at a time. The goal is to learn the physical system (or some properties thereof) from as few copies of it as possible. A learning algorithm with quantum memory can store many copies of the quantum state, and perform joint quantum processing on the product state, followed by an entangled measurement on the quantum memory. On the other hand, a learning algorithm without quantum memory can only perform measurements on each copy of the quantum state, and learn from the resulting classical measurement data. However, note that the choice of measurement applied to the i -th copy could depend on all previous measurement outcomes.

The standard approach for learning physical systems is quantum state tomography [GLF⁺10, OW17, HHJ⁺17, GKKT20a], where the goal is to learn a full representation of ρ up to ε error in trace distance. However, it is well known [OW17, HHJ⁺17] that quantum state tomography requires exponentially many copies $T = 2^{\Omega(n)}$ for learning algorithms with and without quantum memory. Hence, such an approach is intrinsically inefficient.

Shadow tomography To circumvent the exponential scaling, beginning with the work of [Aar18], many researchers [BKL⁺19, AR19, BO20, HKP20, CYZF20] have considered the task of predicting many properties of the quantum state ρ from very few copies of the unknown state. This task is known as *shadow tomography* or *shadow estimation*. Here, we are given M observables O_1, \dots, O_M satisfying $\|O_i\|_\infty \leq 1$, and the goal is to accurately estimate $\text{tr}(O_i \rho)$ to error ε , for all $i = 1, \dots, M$. When $M < 2^{\Omega(n)}$, this can be solved using much fewer copies than is required for full state tomography. In particular, there is a simple algorithm using $\mathcal{O}(M \log M / \varepsilon^2)$ copies that simply measures each observable independently. Notably, this estimator does not require quantum memory.

However, perhaps surprisingly, there are estimators which can solve this task with a sample complexity which is sublinear in M . The most recent progress is due to [BO20], who showed $\mathcal{O}(n \log(M)^2 / \varepsilon^4)$ copies suffice. This is somewhat unexpected, because the observables O_1, \dots, O_M can be highly incompatible, and due to Heisenberg’s uncertainty principle, one should not be able

to efficiently measure these observables simultaneously. Nevertheless, by performing quantum data analysis on a few copies of ρ , we can circumvent this incompatibility.

These estimators crucially rely on fully entangled measurements on all copies of ρ , and hence require a large quantum memory. An important research direction in shadow tomography is understanding how well a learning algorithm without quantum memory can perform. Indeed, this was posed as an open question in the first paper introducing the problem of shadow tomography [Aar18], and has remained an important question in the field ever since. On the upper bounds side, many papers have developed improved algorithms for special classes of observables [PK19, HMR⁺19, HKP20]. See related work for a more in-depth review. However, when no assumptions are made on the class of observables, $\mathcal{O}(M \log M / \varepsilon^2)$ remains the best known sample complexity. On the lower bounds side, a very recent work of [HKP21b] shows that any learning algorithm without quantum memory requires sample complexity at least $\Omega(M^{1/6} / \varepsilon^2)$ for this problem. However, this still left a large gap between the known upper and lower bounds.

In this work, we completely resolve the sample complexity of shadow tomography without quantum memory up to logarithmic factors, by providing a lower bound which (nearly) matches the best known upper bound. In particular, we show:

Theorem 1.1 (General shadow tomography—informal, see Corollary 5.7 and Theorem 5.5). *For all $M > 1$, there exists a set of M observables so that any learning algorithm without quantum memory that solves the shadow tomography problem for this set of observables requires $T = \tilde{\Theta}(\min(M, 2^n) / \varepsilon^2)$ copies.¹*

We note that this bound is tight up to logarithmic factors: there is an algorithm with sample complexity $\mathcal{O}(\min(M \log M, 2^n \log M) / \varepsilon^2)$ [HKP20] that uses no quantum memory. At a high level, this bound says that in general, algorithms without quantum memory cannot do much better than directly measuring each observable. Up to these logarithmic factors, this fully answers the question posed in [Aar18].

Shadow Tomography with Pauli observables An important and well-studied special case of shadow tomography is the case where the observables are the Pauli matrices, or a subset thereof. This problem naturally arises in the study of quantum chemistry [MRBAG16, HMR⁺19, HKP21a]. Indeed, it is this instance which [HKP21b] used to prove their $\Omega(M^{1/6} / \varepsilon^2)$ lower bound. In particular, they show that if the observables are the set of $M = \mathcal{O}(4^n)$ Pauli matrices, then any algorithm which uses no quantum memory and solves the shadow tomography problem must use at least $\Omega(2^{n/3} / \varepsilon^2)$ copies. In contrast, algorithms with quantum memory can solve this instance using $\mathcal{O}(n)$ copies [HKP21b]. As before, the lower bound against algorithms without quantum memory is not tight: in particular, the best known sample complexity upper bound for such algorithms is $\tilde{\mathcal{O}}(2^n)$ [HKP20].

In this work, we provide a tight bound on the sample complexity of this problem without quantum memory, up to lower order terms:

Theorem 1.2 (Pauli observables – informal, see Corollary 5.9 and Theorem 5.5). *To predict the expectation values of all Pauli observables in an unknown n -qubit state ρ , any learning algorithm without quantum memory requires $T = \Omega(2^n)$ copies of ρ .*

Given the previously developed machinery, the proof of this is almost shockingly simple. Indeed, the entire calculation fits in roughly half a page. This serves as a demonstration of the power of the techniques we develop in this paper; in contrast, the proof of the bound in [HKP21b] is substantially more involved, not to mention much looser.

¹Throughout the paper, we say $g = \tilde{\Theta}(f)$ if $g = \Theta(f \log^c f)$ for some fixed c . We similarly define $\tilde{\mathcal{O}}$ and $\tilde{\Omega}$.

Purity testing We next apply our techniques to the slightly different setting of quantum property testing. We are given copies of an n qubit state ρ , but rather than learn statistics of ρ , the goal is to determine whether or not ρ satisfies some condition, or is far from satisfying it. Some examples include mixedness testing (is ρ the maximally mixed state or far from it?) and its generalization identity testing (is $\rho = \sigma$ for some given state σ or far from it?). In this work, we consider the well-studied problem of *purity testing* [EAO⁺02], where the goal is to distinguish if ρ is rank-1, or if it is far from it.

This problem is natural in its own right, but it is of particular interest for us, because it is a setting where there are exponential separations between algorithms that use $\mathcal{O}(n)$ qubits of memory, and those that do not use quantum memory. In fact, this shows up even for the special case of distinguishing whether ρ is a pure state or if ρ is the maximally mixed state. For this case, there is a simple algorithm based on the SWAP test that uses 2-entangled measurements—that is, measurements on $\rho^{\otimes 2}$ —and succeeds with $\mathcal{O}(1)$ samples. Note that this algorithm only requires n qubits of additional quantum memory to store the single-copy from a previous experiment. Prior work of [ACQ21] implied as a corollary that any such algorithm must use at least $\Omega(2^{2n/7})$ samples, but as with shadow tomography, this bound was not tight.

In this work, we completely resolve this question, up to constant factors. In Theorem 5.13, we show that there is an algorithm that uses unentangled measurements with $\mathcal{O}(2^{n/2})$ samples. And, we give the following matching lower bound:

Theorem 1.3 (Purity testing—informal, see Theorem 5.11). *Any algorithm which can distinguish between ρ being a pure state and ρ being the maximally mixed state without using external quantum memory requires $\Omega(2^{n/2})$ samples.*

1.1.2 Smooth tradeoffs for learning with quantum memory

In Section 6 we revisit the problem of shadow tomography with Pauli observables. While our previously established bounds demonstrate exponential separations when the algorithm is given no quantum memory, it is natural to ask whether any algorithm that achieves exponentially better sample complexities must be using a *large* amount of quantum memory. In particular, is it possible that there is an algorithm that only uses a constant number of qubits of external quantum memory, but which still achieves $\mathcal{O}(1)$ sample complexity? We demonstrate that this is impossible, by demonstrating a lower bound which scales with the number of qubits of quantum memory. Formally, we show the following lower bound even for the simpler task of predicting the absolute value of the expectations of the Pauli observables:

Theorem 1.4 (Shadow tomography with bounded quantum memory). *Any learning algorithm with k qubits of quantum memory requires $T \geq \Omega(2^{(n-k)/3})$ copies of ρ to predict $|\text{tr}(P\rho)|$ for all n -qubit Pauli observables P with at least probability $2/3$.*

This bound is meaningful for $k \leq n$, and we complement this with a simple upper bound which uses 2-entangled measurements (and thus n qubits of additional quantum memory to store a single-copy of ρ from the previous experiment) that solves this problem with $\mathcal{O}(n)$ copies. We remark that this still leaves a large part of the landscape open, and we believe it is an interesting open question to fully characterize the sample complexity of this problem as a function of k . However, this result is the first step towards understanding the tradeoffs between sample complexity and quantum memory in a more fine-grained fashion.

1.1.3 Learning quantum dynamics

We now turn our attention to the problem of learning the dynamics of a quantum system. Here, we consider an unknown physical process whose properties we wish to understand. We may perform experiments by interacting with this physical process, and the goal is to learn the desired property of the process with as few experiments as possible. More formally, there is an unknown quantum channel \mathcal{C} over n qubit states. We interact with this channel by passing a quantum state through it and measuring the resulting state. Algorithms with quantum memory may store some quantum information about the result, and subsequent interactions may be entangled with previous outcomes. In contrast, algorithms without quantum memory can only store classical information about the output, although as before, interactions can be chosen adaptively based on the outcomes of previous experiments. In both settings, the goal is to learn some property of \mathcal{C} while minimizing the number of experiments.

A very similar setting was considered in [ACQ21], however, with one key qualitative difference: in [ACQ21], they only consider algorithms without quantum memory that prepare n -qubit states ρ , pass them through the channel \mathcal{C} , and measure the resulting state. For this restricted class of algorithms, they prove exponential separations between the power of algorithms with and without quantum memory.

In general however, an algorithm (even without quantum memory) could also prepare a state with $m > n$ qubits, pass part of it through \mathcal{C} , and perform some arbitrary measurement on the result. Unfortunately, the proof technique in [ACQ21] breaks down even when $m = n + 1$. In particular, the addition of one extra ancilla causes the bound to grow exponentially in the number of experiments. In this work, we generalize and strengthen the lower bounds of [ACQ21] by removing this restriction on ancilla qubits, as well as quantitatively improving some of their bounds. As a result, we demonstrate the first exponential separations for these tasks between *general* algorithms with and without quantum memory.

Distinguishing the completely depolarizing channel The first task we consider is testing whether or not \mathcal{C} is the completely depolarizing channel, that is, the channel which sends every state to the maximally mixed state. Understanding the depolarization of a given quantum channel is a well-studied problem, and is an important component of quantum error mitigation, see e.g. [TBG17] and references within.

We consider the following simple case, previously studied in [BCHJ⁺21, ACQ21]: distinguish between the case where \mathcal{C} is the completely depolarizing channel, and the case where \mathcal{C} is a random unitary channel, that is, $\mathcal{C}[\rho] = U\rho U^\dagger$ for some Haar random unitary matrix U . Even for this very simple case, we demonstrate that algorithms without quantum memory require exponentially more copies of \mathcal{C} than those with quantum memory. Specifically, we show:

Theorem 1.5 (Depolarizing vs. unitary channel—informal, see Theorem 7.9). *Any algorithm without quantum memory that can distinguish between \mathcal{C} being the completely depolarizing channel and \mathcal{C} being a Haar random unitary channel requires $\Omega(2^{n/3})$ experiments.*

As mentioned earlier, this quantitatively improves the lower bound given in [ACQ21], and also applies against all algorithms without quantum memory; in contrast, the lower bound of [ACQ21] only applied to algorithms that did not use ancilla qubits. Combining this bound with the upper bound in [ACQ21], who demonstrated an algorithm that uses quantum memory that only requires $\mathcal{O}(1)$ interactions with \mathcal{C} , yields the claimed exponential separation.

Distinguishing time-reversal symmetries We next turn our attention to the problem of distinguishing between different types of symmetries in a physical evolution. Determining symmetries is one of the most fundamental problems in experimental physics. We will focus on the case of global symmetries, and in particular time-reversal symmetries. Our question will be: does a unitary manifest time-reversal symmetry, and if so what kind?

More precisely, suppose we are given a quantum channel \mathcal{C} , and we are promised it is of the form $\mathcal{C}[\rho] = A\rho A^\dagger$, where A is drawn from the Haar measure of either the class of unitary matrices, orthogonal matrices, or symplectic matrices. Ordinary unitary matrices do not possess time-reversal symmetry, whereas orthogonal matrices and symplectic matrices realize distinct manifestations of time-reversal symmetry (i.e., orthogonal matrices are equal to their own transpose, and symplectic matrices are equal to their own symplectic transpose). The goal is to distinguish which case we are in, with as few experiments as possible. For this problem, there is an algorithm that uses quantum memory that succeeds with $\mathcal{O}(1)$ experiments, however, [ACQ21] previously showed that for the restricted class of algorithms without quantum memory they considered, $\Omega(2^{2n/7})$ experiments were necessary. We generalize this bound to the class of all algorithms without quantum memory:

Theorem 1.6 (Distinguishing time-reversal symmetries—informal, see Theorems 7.9, 7.10, and 7.11). *Any algorithm without quantum memory that can distinguish between time-reversal symmetries with high probability requires $\Omega(2^{2n/7})$ experiments.*

2 Technical Overview

As is common in the property testing literature, the starting point for our lower bounds is Le Cam’s two-point method (see Lemma 5.3). We would like to exhibit a simple “null hypothesis” and a “mixture of alternatives” and argue that distinguishing between these two scenarios with nontrivial advantage requires a large number of quantum measurements.

As a first running example, consider the task of shadow tomography with Pauli observables, where the goal is to estimate the expectation value of all Pauli observables to ε -error. The corresponding distinguishing task will be to decide whether the unknown state is the maximally mixed state ρ_{mm} or otherwise equal to $\rho_P \triangleq \frac{1}{2^n}(\mathbb{1} + \varepsilon P)$ for a uniformly random signed Pauli matrix $P \neq \mathbb{1}$. Given a learning algorithm that makes T measurements, let q_0 and $q_1 = \mathbb{E}_P[q_1^P]$ denote the (classical) distributions over measurement outcomes induced by the algorithm in these two situations, where q_1^P denotes the distribution over outcomes when the unknown state is ρ_P . To prove the $T = \Omega(2^n/\varepsilon^2)$ lower bound in Theorem 1.2, it suffices to show that for $T = o(2^n/\varepsilon^2)$, $d_{\text{TV}}(q_0, \mathbb{E}_P[q_1^P]) = o(1)$ (see Section 4.3).

We emphasize that while this general approach of proving statistical lower bounds by showing estimates on the total variation between a single “simple” distribution and a mixture is standard, the general technique we use for proving these estimates is a departure from standard approaches. Indeed, whereas the latter typically first pass to chi-squared or KL divergence [Pan08, ACBFS02, BG18, BCL20, IS12], we work directly with total variation. We remark that this is in some sense necessary: for our Pauli observables lower bound for instance, passing to chi-squared divergence in the first step will result in a lower bound of at most $T = \Omega(\text{poly}(n, 1/\varepsilon))$.

One of the general techniques that we employ to bound total variation is to reduce to showing **one-sided likelihood ratio bounds**, that is, to go through the following elementary fact (see also Lemma 5.4).

Fact 2.1. *Let $0 \leq \delta < 1$. Given probability distributions p, q on a finite domain Ω , if $p(x)/q(x) > 1 - \delta$ for all $x \in \Omega$, then $d_{\text{TV}}(p, q) \leq \delta$.*

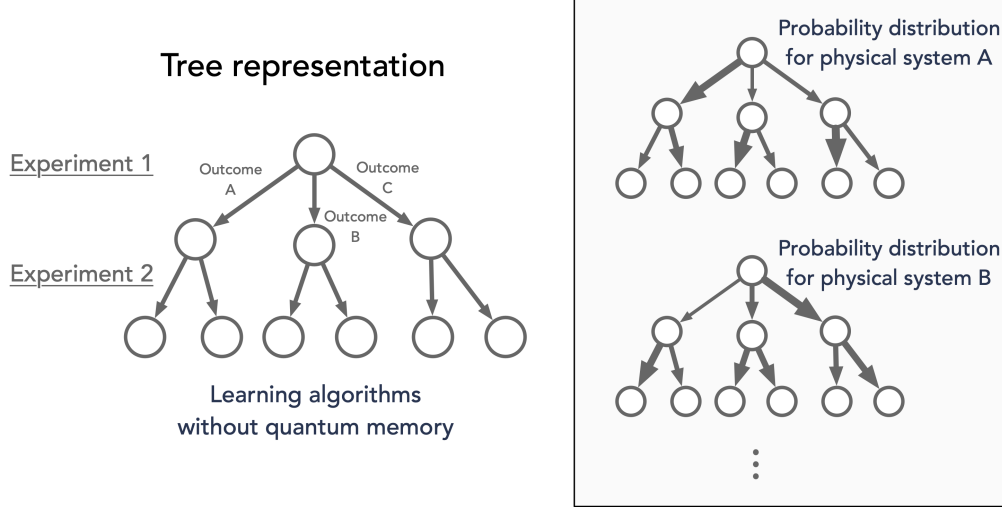


Figure 1: Illustration of the tree representation for learning algorithms without quantum memory. When the unknown physical system (or physical process) is different, the probability to traverse each edge will be different. Learning succeeds if for distinct physical systems, the distributions over leaf nodes are sufficiently different.

Fact 2.1 tells us that to bound $d_{\text{TV}}(q_0, \mathbb{E}_P[q_1^P])$, it suffices to show a *uniform* lower bound on the likelihood ratio

$$\mathbb{E}_P[q_1^P(x)]/q_0(x). \quad (1)$$

In words, it suffices to show that no sequence of measurement outcomes from applying the learning algorithm to samples from the mixture of alternatives is too unlikely compared to applying it to samples from the null hypothesis. While such an estimate need not hold in general, e.g. for our channel learning and bounded memory lower bounds which call for more involved techniques, it turns out to hold for all of our lower bounds on learning quantum states. Furthermore, our lower bounds on learning quantum states all match known upper bounds up to a logarithmic factor. Indeed, the fact that such an approach can yield tight or nearly tight lower bounds for these problems is quite surprising.

2.1 Tree representation

A primary technical hurdle to overcome which does not manifest in analogous classical settings is the fact that the quantum experiments that the learner runs may be chosen *adaptively* based on the outcomes of earlier experiments. Such algorithms are naturally represented as a tree [BCL20, ACQ21, HKP21b]. Formally, we associate to any such algorithm a rooted tree \mathcal{T} in the following way. The root r of the tree corresponds to the initial state of the algorithm. At any node u , the algorithm performs an experiment. Concretely, if the goal is to learn a quantum state, it makes a POVM measurement on a fresh copy of the state; alternatively, if the goal is to learn a quantum channel, it prepares an input quantum state, passes it through the channel, and makes a POVM measurement on the output. After the experiment, the state of the algorithm moves to a child node of u depending on the experimental outcome. In this way, each node encodes all experimental outcomes that have been seen so far; correspondingly, the T -th layer of the rooted tree corresponds to all possible states of the algorithm after T experiments.

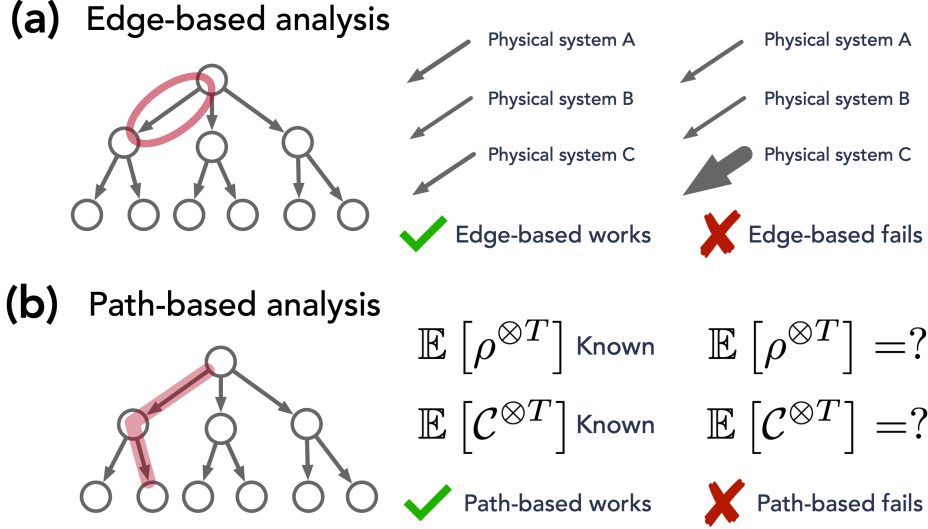


Figure 2: Illustration of edge-based and path-based analysis proving lower bounds for learning algorithms without quantum memory. Edge-based analysis works when the probability to traverse every edge does not differ too much across different unknown physical system (or process). Path-based analysis works when we have an adequate characterization for the higher moments for the distribution over unknown physical system (or process).

The tree structure naturally encodes the adaptivity of the learning algorithm, as the choice of experiment that the algorithm performs at a given node can depend on all prior experimental outcomes. Bounding the aforementioned total variation distance thus amounts to bounding the total variation distance between the induced distribution on leaves under the null hypothesis versus under the mixture of alternatives.

With this tree formalism in place, we can now describe our different techniques for bounding total variation, which roughly fall into two categories:

1. **Edge-based:** Bound the information the learner can gain from traversing *any* edge in the tree.
2. **Path-based:** Utilize multi-linear structure in the tree and high-order moments of the unknown states or channels to bound the information gain for traversing an entire path jointly.

2.2 Edge-based analysis

We formalize the notion of “information” gained from any given edge as follows. As an example, consider again the task of shadow tomography for Pauli observables, and suppose for simplicity of exposition that at every node u , the algorithm performs some projective measurement $\{|\psi_s^u\rangle\langle\psi_s^u|\}_s$, where every s corresponds to a child of u .² For each edge $e_{u,s}$, we can consider the following notion of **average discrepancy** along an edge, namely the squared difference between the probability of traversing $e_{u,s}$ when the underlying state is ρ_{mm} versus when it is ρ_P , averaged over the mixture of alternatives P :

$$\mathbb{E}_P[(\langle\psi_s^u|\rho_{\text{mm}}|\psi_s^u\rangle - \langle\psi_s^u|\rho_P|\psi_s^u\rangle)^2], \quad (2)$$

²As we will show, the proceeding discussion extends to general POVMs.

where $\langle \psi_s^u | \rho_{(\cdot)} | \psi_s^u \rangle$ is the probability for traversing edge $e_{u,s}$ when the unknown state is $\rho_{(\cdot)}$.

Given bounds on the average discrepancy in Eq. (2), we can leverage a convexity argument to obtain a one-sided likelihood ratio bound (see the calculation beginning at Eq.). The one-sided likelihood ratio bound would then give us the lower bound on T through Fact 2.1. A similar convexity argument appeared implicitly in [BCL20]. We generalize this argument, and in doing so, elucidate a new quantity $\delta(O_1, \dots, O_M)$ (see Eq. (79)) for a collection of any M observables O_1, \dots, O_M that, loosely speaking, measures how much information an algorithm without quantum memory can learn from a single measurement. For instance, in the Pauli setting, we can upper bound δ by $\mathcal{O}(1/2^n)$ (see Lemma 5.8), yielding our near-optimal lower bound against algorithms without quantum memory for shadow tomography with Pauli observables. For shadow tomography with general observables, we obtain an analogous result, by applying the bound to a collection of Haar-random observables.

At this juncture, it is instructive to underscore how crucial it is that *one-sided* bounds on the likelihood ratio suffice to bound total variation. Indeed, even though a bound on (2) implies that traversing any given edge is roughly equally likely on average under the null hypothesis as it is under the mixture of alternatives, it *does not* imply that traversing any given *path* from root to leaf ℓ is roughly equally likely. That is, the likelihood ratio $\mathbb{E}_P[p^{\rho_P}(\ell)]/p^{\rho^{\text{mm}}}(\ell)$ could be quite large; in fact it could be much larger than 1, because the discrepancies introduced by the individual edges along the path could easily compound. For instance, if in every experiment the learner simply measured in the computational basis, then for the root-to-leaf path corresponding to observing $|0\rangle$ after every measurement, (1) would be of order $(1+\varepsilon)^T$. If we simply upper bounded total variation by the average absolute difference between (1) and 1, we would at best obtain a $T = \Omega(n/\varepsilon)$ sample complexity lower bound.

Shadow tomography with k -qubit memory. For proving lower bounds against learning algorithms with k qubits of memory, we can employ a similar idea. Whereas previously, we merely associated to each node the probability of reaching it, we now associate to it a $2^k \times 2^k$ positive-semidefinite (PSD) matrix representing the unnormalized quantum state of the k -qubit memory. Each edge can then be regarded as a completely-positive linear map on PSD matrices. We can define and upper bound a notion of average discrepancy along an edge generalizing (2) (see Lemma 6.3).

The main difference is that the aforementioned convexity argument for passing from an average discrepancy bound to a likelihood ratio bound no longer applies to give a uniform lower bound on the likelihood ratio. Instead, we utilize a **pruning argument**. By Markov's, a bound on the average discrepancy along an edge implies that the discrepancy for a random Pauli P along that edge is small with high probability over the choice of P (see Definition 6.4 and Fact 6.5). In particular, for any leaf ℓ in the tree, the fraction of Paulis for which the discrepancy is small for every edge in the path from root to ℓ is also large by a union bound. So for any leaf ℓ , if we decompose the likelihood ratio (1) into the contribution from the Paulis with small discrepancy and large discrepancy. Those of the latter type are few enough that we can naively prune them from consideration (Lemma 6.6). The contribution from those of the former type, averaged across leaves, can be related to the total variation distance at depth $T - 1$. Proceeding by induction we can obtain the desired sample complexity lower bound.

2.3 Path-based analysis

From average discrepancy to multi-linear structure. The edge-based analysis fails when there is an extensive number of edges where the average discrepancy is sizable. As a warmup example, consider the problem of purity testing, where the null hypothesis is that ρ is maximally mixed,

and the mixture of alternatives is that ρ is a random pure state. As above, suppose for simplicity of exposition that the quantum experiments the learner runs are all projective measurements $\{|\psi_s^u\rangle\langle\psi_s^u|\}_s$. Then for all edges $e_{u,s}$ the average discrepancy $\mathbb{E}_{|\phi\rangle}[(\langle\psi_s^u|\rho_{\text{mm}}|\psi_s^u\rangle - \langle\psi_s^u|\phi\rangle\langle\phi|\psi_s^u\rangle)^2]$ between the completely mixed state ρ_{mm} and a random pure state $|\phi\rangle\langle\phi|$ is of order $\mathcal{O}(1/2^{2n})$. Using the aforementioned convexity argument, this would only yield a trivial lower bound of $\Omega(1)$.

In order to establish a nontrivial lower bound, we consider an analysis that better utilize structure in the tree representation of the learning algorithm. The central idea is to exploit the **multi-linear structure** inherent in the tree representation. We illustrate this idea using the purity testing problem. For a given leaf ℓ , suppose the edges on the path from root to ℓ are $e_{u_0=r,s_0}, e_{u_1,s_1}, \dots, e_{u_{T-1},s_{T-1}}$, where $e_{u_{T-1},s_{T-1}}$ connects u_{T-1} and ℓ . The probability associated with the leaf ℓ under an unknown quantum state ρ is given by

$$p^\rho(\ell) = \prod_{t=0}^{T-1} (\langle\psi_{s_t}^{u_t}|\rho|\psi_{s_t}^{u_t}\rangle) = \text{tr} \left(\left(\bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right) \rho^{\otimes T} \right). \quad (3)$$

Under the tree representation, the measurement outcomes along the path yield a tensor product of pure states $\bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}|$ that depends on the learning algorithm. The key point is that the probability (3) of traversing this path is linear in $|\psi_s^u\rangle\langle\psi_s^u|$ for every edge $e_{u,s}$ on the path, and also linear in $\rho^{\otimes T}$.

How does $p^\rho(\ell)$ behave under averaging of ρ across a mixture of alternatives? In our purity testing and channel learning lower bounds, the mixture of alternatives is an average over the Haar measure of some Lie group, so we can readily compute $\mathbb{E}_\rho[p^\rho(\ell)]$ by computing the high-order moment $\mathbb{E}_\rho[\rho^{\otimes T}]$.

Purity testing. We first instantiate this for purity testing. When ρ is a random pure state $|\phi\rangle\langle\phi|$, then it is well-known [Chr06, Har13] that $\mathbb{E}[\rho^{\otimes T}]$ is proportional to the projector Π onto the symmetric subspace in $(\mathbb{C}^{2^n \times 2^n})^{\otimes T}$ (see Lemma 4.12). Specifically, the probability for a leaf ℓ is exactly $\prod_{t=0}^{T-1} (1 - t/2^n)^{-1} \cdot \text{tr} \left(\Pi \bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right)$. The key inequality we show is that

$$\text{tr} \left(\Pi \bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right) \geq 1, \quad (4)$$

i.e. the squared norm of the projection of $\bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}|$ onto the symmetric subspace, is always at least 1 (see Lemma 5.12). The upshot is that we obtain a uniform one-sided lower bound on the likelihood ratio of $\prod_{t=0}^{T-1} (1 - t/2^n)^{-1}$, and for $T = o(2^{n/2})$ this quantity is $1 - o(1)$ as desired.

To get intuition for this one-sided lower bound, imagine the projective measurements $\{|\psi_s^u\rangle\langle\psi_s^u|\}$ were all in the same basis. Then the sequence of outcomes which would be most unlikely for the mixture of alternatives (random pure state) relative to the null hypothesis (maximally mixed state) would be one in which all $|\psi_s^u\rangle$ along the path were mutually orthogonal. Note that in such a case, we actually have $\text{tr} \left(\Pi \bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right) = 1$. Moreover, the quantity $\prod_{t=0}^{T-1} (1 - t/2^n)$ is the probability that all T outcomes from measuring $\rho = \rho_{\text{mm}}$ are distinct. Equivalently, this is the probability of seeing a collision among T draws from the uniform distribution $[2^n]$, and this is exactly why we obtain an $\Omega(2^{n/2})$ lower bound for purity testing. We remark that extending this intuition to a formal proof of Eq. (4) is nontrivial and requires a careful tensor analysis that crucially exploits the multi-linear structure afforded by the tree representation.

Extending to channel learning. The multi-linear structure also plays a crucial role for establishing lower bounds for channel learning tasks. Consider the task of distinguishing between whether the unknown channel \mathcal{C} is a completely depolarizing channel \mathcal{D} or a random channel from some distribution. Each edge $e_{u,s}$ in the tree representation of an algorithm (without quantum memory) for learning quantum channels is characterized by an $(n + n')$ -qubit input state $|\phi^u\rangle$ and an $(n + n')$ -qubit measurement outcome $|\psi_s^u\rangle$. Again for simplicity, we only consider projective measurements in the technical overview.

Similar to purity testing, we consider a sequence of edges $\{e_{u_t, s_t}\}$ on a path from root to leaf ℓ . The probability that we end up at ℓ when we employ the learning algorithm on an unknown quantum channel \mathcal{C} can be shown to be

$$p^{\mathcal{C}}(\ell) = \prod_{t=0}^{T-1} \langle \psi_{s_t}^{u_t} | (\mathcal{C} \otimes \mathbb{1}) (|\phi^{u_t}\rangle\langle\phi^{u_t}|) | \psi_{s_t}^{u_t} \rangle = \text{tr} \left(\left(\bigotimes_{t=0}^{T-1} |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right) \left(\bigotimes_{t=0}^{T-1} (\mathcal{C} \otimes \mathbb{1}) (|\phi^{u_t}\rangle\langle\phi^{u_t}|) \right) \right). \quad (5)$$

Hence, the average probability $\mathbb{E}_{\mathcal{C}}[p^{\mathcal{C}}(\ell)]$ would be a multi-linear function in $\mathbb{E}_{\mathcal{C}}[\bigotimes_{t=0}^{T-1} (\mathcal{C} \otimes \mathbb{1})]$, the T -th moment of the unknown quantum channel \mathcal{C} , and elements $|\phi^u\rangle, |\psi_s^u\rangle$ in the algorithm.

The main complication in channel learning tasks is that it becomes much more challenging to obtain a uniform one-sided likelihood ratio bound. Instead, we explicitly expand the T -th moment of the unknown quantum channel \mathcal{C} into a linear combination of basic tensors, e.g. permutation operators, that arise from a suitable Weingarten expansion of the integral over \mathcal{C} . The total variation distance can be upper bounded by a sum over absolute values of multi-linear functions,

$$\sum_{\ell: \text{leaves}} |p^{\mathcal{D}}(\ell) - \mathbb{E}_{\mathcal{C}}[p^{\mathcal{C}}(\ell)]| \leq \sum_{\{u_t, s_t\}_{t=0}^{T-1}} \sum_A \left| \text{tr} \left(\underbrace{A \bigotimes_{t=0}^{T-1} (|\phi^{u_t}\rangle\langle\phi^{u_t}| \otimes |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}|)}_{\text{Multi-linear function}} \right) \right|. \quad (6)$$

The sum over A depends on the T -th moment of the unknown channel $\mathbb{E}_{\mathcal{C}}[(\mathcal{C} \otimes \mathbb{1})^{\otimes T}]$. For each multi-linear function, the interactions between the tensor components $|\phi^{u_t}\rangle, |\psi_{s_t}^{u_t}\rangle$ can be rather complex. Hence a central tool in our analysis is to utilize tensor network diagrams [Orú19] to bound and disentangle the relations between each components.

The goal of the tensor network manipulation is to upper bound the total variation distance by a sum over multi-linear functions without the absolute values,

$$\sum_{\{u_t, s_t\}_{t=0}^{T-1}} \sum_B \text{tr} \left(B \bigotimes_{t=0}^{T-1} (|\phi^{u_t}\rangle\langle\phi^{u_t}| \otimes |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}|) \right). \quad (7)$$

Then we can utilize the resolution of identity $\sum_s |\psi_s^u\rangle\langle\psi_s^u| = \mathbb{1}$ at the bottom layer of the tree to collapse the summation over all leaves (nodes at depth T) to summation over nodes at depth $T-1$. In other words, we will upper bound $\text{tr}(B(X \otimes |\phi^{u_{T-1}}\rangle\langle\phi^{u_{T-1}}| \otimes \mathbb{1})) \leq \text{tr}(B'X)$ for some tensor B' , which lets us upper bound (7) by

$$\sum_{\{u_t, s_t\}_{t=0}^{T-2}} \sum_{B'} \text{tr} \left(B' \bigotimes_{t=0}^{T-2} (|\phi^{u_t}\rangle\langle\phi^{u_t}| \otimes |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}|) \right). \quad (8)$$

We inductively perform this collapsing procedure to reduce the summation over all paths on the tree to a single value and conclude the lower bound proof.

3 Related Work

The field of learning and testing properties of quantum systems is broad and a full review of the literature is out of scope for this paper. For conciseness, here we only review the most relevant lines of work. See e.g. [MdW13] for a more complete survey.

Learning properties of quantum states When it comes to learning properties of quantum states, arguably the most canonical problem is quantum tomography, which is the problem of learning the unknown quantum state ρ to high accuracy, usually in trace norm or fidelity. See e.g. [Hra97, GLF⁺10, BK10, BCG13] and references within. It was not until recently that the sample complexity of this problem was fully characterized [HHJ⁺17, OW16]. These papers demonstrated that $\Theta(2^{2n})$ samples are necessary and sufficient for this problem. In particular, the sample complexity is exponential in n .

As a way around this barrier, [Aar18] proposed the shadow tomography problem. Ever since, there has been an active line of work improving the sample complexity upper bounds for this problem [BKL⁺19, AR19, HKP20, BO20, HKP21b]. However, the algorithms presented in these papers largely require heavily entangled measurements, and thus a great deal of quantum memory. Some notable exceptions are [HKP20, HKP21b], who gave algorithms for shadow tomography that do not use quantum memory that match the lower bound we achieve in the general case, and improved algorithms for shadow tomography in special cases. The special case that we consider in this paper where the observables are Pauli matrices is also of independent interest, as it is a useful subroutine in many quantum computing applications, see discussions in [MRBAG16, KMT⁺17, PK19, YVI19, GAD⁺19, HMR⁺19, CW20, CvSW⁺20, HKP21a, HKT⁺21].

In quantum property testing, besides purity testing, there have been a number of well-studied problems, including quantum state discrimination [Che00, ANSV08, BC09], mixedness testing [OW15, BCL20] and more generally spectrum testing [OW15] and state certification [BOW19, CLO21]. However, for these problems, while in some settings there are separations between the power of algorithms with and without quantum memory, these settings inherently limit the separations to be at most polynomial, rather than exponential.

Quantum memory tradeoffs for learning Understanding the power of algorithms without quantum memory has been posed as an open question in a number of learning theoretic settings, such as shadow tomography [Aar18], as well as spectrum testing and general state tomography [Wri16]. However, until recently, lower bounds against the power of algorithms without quantum memory usually only applied to the non-adaptive setting, where the algorithm had to commit to the set of measurements ahead of time, see e.g. [HHJ⁺17, CLO21]. The first work which demonstrated a separation against general algorithms was [BCL20], who showed a polynomial gap for mixedness testing and quantum identity testing. Subsequently, the aforementioned works of [HKP20, ACQ21] demonstrated exponential separations for shadow tomography and channel learning. In many ways this work can be thought of as a synthesis of the ideas in [BCL20] with the ones developed in [HKP20, ACQ21] to both tighten and generalize the bounds obtained by the latter. We also note that memory tradeoffs have also been demonstrated for classical learning problems, e.g. [SVW16, Raz18, Raz17, KRT17a, MM17, MM18, SSV19]. It would be interesting to see if there are any technical connections between these two lines of inquiry.

Learning quantum dynamics Similar to learning quantum states, the task of learning a full description of a quantum process/channel is known as quantum process tomography [MRL08].

Due to the exceedingly high sample complexity, most of the recent works have focused on learning restricted classes of dynamics. A large class of algorithms assume the dynamics to be governed by a time-independent Hamiltonian and the goal is to reconstruct the Hamiltonian [WGFC14, WPS⁺17, EHF19]. Another class of quantum channels that have been subject to active study is Pauli channels [HFW20, FW20, FO21], which are useful for characterizing noise in a quantum device. More recently, [HKP21b, ACQ21, RYCS21, CZSJ21] studied the task of learning properties of quantum channels using information-theoretic bounds.

4 Preliminaries

Given a (possibly unnormalized) state σ on $m \geq k$ qubits, let $\text{tr}_{>k}(\sigma)$ denote the state on k qubits obtained by tracing out the last $m - k$ qubits. We denote the maximally mixed state on n qubits by $\rho_{\text{mm}} \in \mathbb{C}^{2^n \times 2^n}$. Given signed Pauli matrix $P \in \mathbb{C}^{2^n \times 2^n}$, define $\rho_P \triangleq (\mathbb{1} + P)/2^n$. We will also use the small-o notation $f(n) = o(g(n))$ if for all $c > 0$ there exists a $k > 0$, such that $0 \leq f(n) < cg(n), \forall n \geq k$. A fact we will often use in the proof is the following: if $g(n) = o(h(n))$ implies $f(n) = o(1)$, then $f(n) = \Omega(1)$ implies $g(n) = \Omega(h(n))$. Given a matrix M , we will use M^\dagger to denote conjugate transpose and M^t to denote transpose. Given probability distributions p, q over some domain Ω , define the *total variation distance* $d_{\text{TV}}(p, q) \triangleq \sup_{S \subseteq \Omega} p(S) - q(S)$.

4.1 Tail bounds

We also recall the following properties for subexponential random variables. A mean-zero random variable Z is λ -subexponential if it satisfies the tail bound $\Pr[|Z| > t] \leq 2e^{-2t/\lambda}$.

Fact 4.1 (See e.g. Lemma 2.7.6 in [Ver18]). *If X is sub-Gaussian with variance proxy σ^2 , then $Z \triangleq X^2 - \mathbb{E}[X^2]$ is $O(\sigma^2)$ -subexponential.*

Fact 4.2 (Bernstein's, see e.g. Corollary 2.8.3 in [Ver18]). *If Z_1, \dots, Z_m are independent, mean-zero, λ -subexponential random variables, then for $Z \triangleq \frac{1}{m} \sum_i Z_i$ and any $t > 0$,*

$$\Pr[|Z| > t] \leq \exp \left(-\frac{m}{2} \left(\min \left(\frac{t^2}{\lambda^2}, \frac{t}{\lambda} \right) \right) \right). \quad (9)$$

We will also use concentration of measure for the Haar measures on the orthogonal and unitary groups:

Lemma 4.3 (See e.g. [AGZ09], Corollary 4.4.28). *For $G = O(d), U(d)$, let $f : G \rightarrow \mathbb{R}$ be L -Lipschitz with respect to the Frobenius norm. There is an absolute constant $c > 0$ such that for x sampled from the Haar measure on G , $\Pr[|f(x) - \mathbb{E}[f(x)]| > c \cdot L \sqrt{\log(1/\delta)/d}] \leq \delta$.*

4.2 Basic results in quantum information theory

In the following, we present several standard definitions in quantum information theory. We also provide the basic lemmas that illustrate properties of these quantum objects.

Definition 4.4 (Quantum channel). *A quantum channel \mathcal{C} from n -qubit to m -qubit is a linear operator from $\mathbb{H}^{2^n \times 2^n}$ to $\mathbb{H}^{2^m \times 2^m}$ given by*

$$\mathcal{C}(X) = \sum_i K_i X K_i^\dagger, \quad (10)$$

where $K_i \in \mathbb{H}^{2^m \times 2^n}$ satisfies $\sum_i K_i^\dagger K_i = \mathbb{1}$. When $n = m$, we let \mathcal{I} denote the identity channel, that is, the channel that acts trivially on $\mathbb{H}^{2^n \times 2^n}$.

Definition 4.5 (POVMs). A positive operator-valued measure (POVM) on n -qubit states is given by positive-semidefinite matrices $\{F_s\}_s$, where the set of matrices $F_s \in \mathbb{C}^{2^n \times 2^n}$ satisfies $\sum_s F_s = \mathbb{1}$. The probability for obtaining classical outcome s when we measure the density matrix $\rho \in \mathbb{H}^{2^n \times 2^n}$ is $\text{tr}(F_s \rho)$.

In Section 6, we will need to reason about how states transform upon being measured; we summarize the formalism for this in the following definition:

Definition 4.6 (Post-measurement states). Given a POVM on n -qubit states $\{F_s\}_s$ where each POVM element has some Cholesky decomposition $F_s = M_s^\dagger M_s$ for $M_s \in \mathbb{C}^{2^{n+m} \times 2^n}$, the post-measurement quantum state upon measuring ρ with this POVM and observing outcome s is given by

$$\frac{M_s \rho M_s^\dagger}{\text{tr}(M_s^\dagger M_s \rho)} \in \mathbb{H}^{2^{n+m} \times 2^{n+m}}. \quad (11)$$

We also define a restricted version of POVM where each POVM element is a rank-1 matrix.

Definition 4.7 (Rank-1 POVM). A rank-1 positive operator-valued measure (POVM) on n -qubit state is given by $\{\sqrt{w_s} 2^n |\psi_s\rangle\langle\psi_s|\}_s$, where $w_s \geq 0$, $w_s \in \mathbb{R}$ and pure states $\{|\psi_s\rangle\}_s$ satisfy

$$\sum_s w_s 2^n |\psi_s\rangle\langle\psi_s| = \mathbb{1}. \quad (12)$$

The above normalization condition implies $\sum_s w_s = 1$. The probability for obtaining the classical outcome s when we measure the density matrix $\rho \in \mathbb{H}^{2^n \times 2^n}$ is $w_s 2^n \langle\psi_s|\rho|\psi_s\rangle$. And the post-measurement quantum state is given by $|\psi_s\rangle\langle\psi_s|$.

A nice property of rank-1 POVM is that it can be used to simulate arbitrary POVM. Hence, rank-1 POVM is as powerful as any POVM from an information-theoretic perspective. The following folklore lemma illustrates this fact. We provide a proof of this lemma for completeness.

Lemma 4.8 (Simulating POVM with Rank-1 POVM). When we only consider the classical outcome of the POVM measurement and neglect the post-measurement quantum state, then an POVM $\{F_s\}_s$ on n -qubit states can be simulated by a rank-1 POVM on n -qubit states with some post-processing.

Proof. Let F_s have Cholesky decomposition $M_s^\dagger M_s$. Consider the diagonalization

$$M_s^\dagger M_s = \sum_p \tilde{w}_{sp} |\psi_{sp}\rangle\langle\psi_{sp}| = \sum_p w_{sp} 2^n |\psi_{sp}\rangle\langle\psi_{sp}|, \quad (13)$$

where $w_{sp} = \tilde{w}_{sp}/2^n$. Because $M_s^\dagger M_s$ is positive-semidefinite, we have $w_{sp} \geq 0$. $\{\sqrt{w_{sp} 2^n} |\psi_{sp}\rangle\langle\psi_{sp}|\}_{sp}$ forms a rank-1 POVM because $\sum_{sp} w_{sp} 2^n |\psi_{sp}\rangle\langle\psi_{sp}| = \sum_s M_s^\dagger M_s = \mathbb{1}$. To simulate the POVM $\{F_s\}_s$, we measure the rank-1 POVM on ρ , which gives the classical outcome sp with probability

$$w_{sp} 2^n \langle\psi_{sp}|\rho|\psi_{sp}\rangle = \text{tr}(w_{sp} 2^n |\psi_{sp}\rangle\langle\psi_{sp}|\rho), \quad (14)$$

then output s . Hence, we output classical outcome s with probability

$$\sum_p \text{tr}(w_{sp} 2^n |\psi_{sp}\rangle\langle\psi_{sp}|\rho) = \text{tr}(M_s^\dagger M_s \rho). \quad (15)$$

Thus the rank-1 POVM with some post-processing simulates the POVM $\{M_s\}_s$. \square

Definition 4.9 (Pauli observables). We define an n -qubit Pauli observable P to be a tensor product of n operators from the set $\left\{ I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$.

Lemma 4.10. The sum of the tensor product of two Pauli observables is given by

$$\sum_{P \in \{I, X, Y, Z\}^{\otimes n}} P \otimes P = 2^n \text{SWAP}_n, \quad (16)$$

where SWAP_n is the swap operator on two copies of n -qubit state.

Proof. It is easy to check the following equality,

$$I \otimes I + X \otimes X + Y \otimes Y + Z \otimes Z = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} = 2\text{SWAP}_1. \quad (17)$$

Then note that $\sum_{P \in \{I, X, Y, Z\}^{\otimes n}} P \otimes P = \left(\sum_{\sigma \in \{I, X, Y, Z\}} \sigma \otimes \sigma \right)^{\otimes n} = 2^n \text{SWAP}_1^{\otimes n} = 2^n \text{SWAP}_n$. \square

In the following, we give basic definition for permutation operators over a tensor product space. In particular, we will use the same notation π for an element in the symmetric group as well as the permutation operator over the tensor product space.

Definition 4.11 (Permutation operators). For $T > 0$, we consider S_T to be the symmetric group of degree T . For any permutation $\pi \in S_T$, we consider the action of π on a tensor product of T states to be given by the following

$$\pi(|\psi_1\rangle \otimes \dots \otimes |\psi_T\rangle) = |\psi_{\pi^{-1}(1)}\rangle \otimes \dots \otimes |\psi_{\pi^{-1}(T)}\rangle, \forall \psi_1, \dots, \psi_T. \quad (18)$$

We linearly extend the action of π to any element in the tensor product space.

Lemma 4.12 (Haar integration over states, see e.g. [Har13]). Consider the uniform (Haar) measure over n -qubit pure states $|\psi\rangle$, then

$$\mathbb{E}_{|\psi\rangle} [|\psi\rangle\langle\psi|^{\otimes T}] = \binom{2^n + T - 1}{T}^{-1} \sum_{\pi \in S_T} \pi, \quad (19)$$

where π is a permutation operator on a tensor product space of T n -qubit pure states and S_T is the symmetric group of degree T .

4.3 Primer on information-theoretic lower bounds

In this section, we give a brief overview for how to prove information-theoretic lower bounds like the ones in this work.

As discussed in Section 2, we are interested in *distinguishing tasks*. Informally, we have the ability to make some number N of measurements of an unknown object z which is promised to lie in one of two possible sets \mathcal{S}_0 and \mathcal{S}_1 . Based on the sequence of measurement outcomes we obtain, we must decide whether the object lies in \mathcal{S}_0 or \mathcal{S}_1 . In this work we will focus on tasks where \mathcal{S}_0 is a singleton set.

We will not formally define what it means to follow a particular “strategy” for measuring z , as this will be clear in the body of this paper (see e.g. Definitions 5.1, 7.1, and 6.1). In any case, this

is not important for the formalism in the present subsection; the only thing one needs to know here is that any z and measurement strategy induce a probability distribution over *transcripts*, namely sequences of measurement outcomes. And as we are interested in showing lower bounds against arbitrary measurement strategies, in the rest of this subsection we may as well fix an arbitrary such strategy.

Under this strategy, to any z is associated a distribution $p_z^{\leq N}$ over transcripts— when z is the unique element of \mathcal{S}_0 , we will denote this by $p_0^{\leq N}$. To show a lower bound for the distinguishing task, formally we would like to argue that there exists no algorithm \mathcal{A} mapping transcripts T to $\{0, 1\}$ for which $\Pr_{T \sim p_z^{\leq N}}[\mathcal{A}(T) = i] \geq 2/3$ for all $z \in \mathcal{S}_i$ and $i = 0, 1$ (the constant $2/3$ is arbitrary and we could consider any value greater than $1/2$).

To show this, it suffices to consider the following *average-case* version of the original distinguishing task. Let \mathcal{D} be some distribution over \mathcal{S}_1 that we are free to choose, and consider the *hypothesis testing* problem of distinguishing whether a given transcript came from $p_0^{\leq N}$ or $\mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}]$. The following fact is an elementary but fundamental result in binary hypothesis testing:

Fact 4.13. *Given distributions q_0, q_1 over a domain \mathcal{S} , if $d_{TV}(q_0, q_1) < 1/3$, there is no algorithm $\mathcal{A} : \mathcal{S} \rightarrow \{0, 1\}$ for which $\Pr_{x \sim q_i}[\mathcal{A}(x) = i] \geq 2/3$ for both $i = 0, 1$.*

Proof. Let $\mathcal{S}' \subseteq \mathcal{S}$ denote the set of elements x for which $\mathcal{A}(x) = 0$. Then observe that

$$\Pr_{x \sim q_0}[\mathcal{A}(x) = 1] + \Pr_{x \sim q_1}[\mathcal{A}(x) = 0] = 1 - q_0(\mathcal{S}') + q_1(\mathcal{S}') \quad (20)$$

$$\geq 1 - \sup_{\mathcal{S}'' \subseteq \mathcal{S}} |q_0(\mathcal{S}'') - q_1(\mathcal{S}'')| \quad (21)$$

$$= 1 - d_{TV}(q_0, q_1) \geq 2/3, \quad (22)$$

so at least one of the terms on the left-hand side is at least $1/3$. \square

We can use this to show that, in order to prove a lower bound for the original distinguishing task, it suffices to bound $d_{TV}(p_0^{\leq N}, \mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}])$:

Lemma 4.14 (Le Cam's two-point method). *If there exists a distribution \mathcal{D} over \mathcal{S}_1 for which $d_{TV}(p_0^{\leq N}, \mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}]) < 1/3$, there is no algorithm \mathcal{A} which maps transcripts of N measurement outcomes to $\{0, 1\}$ for which $\Pr_{T \sim p_z^{\leq N}}[\mathcal{A}(T) = i] \geq 2/3$ for any $z \in \mathcal{S}_i$ and $i = 0, 1$.*

Proof. Suppose to the contrary that there existed such an algorithm. Let $p_1^{\leq N} \triangleq \mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}]$. Then

$$2/3 \leq \mathbb{E}_{z \sim \mathcal{D}} \left[\Pr_{T \sim p_z^{\leq N}}[\mathcal{A}(T) = i] \right] = \Pr_{T \sim p_1^{\leq N}}[\mathcal{A}(T) = i]. \quad (23)$$

But by Fact 4.13, this would contradict the fact that $d_{TV}(p_0^{\leq N}, \mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}]) < 1/3$. \square

In the distinguishing tasks we consider, the choice of \mathcal{D} will be fairly clear, so the primary technical difficulty for us will be upper bounding the total variation distance between $p_0^{\leq N}$ and $\mathbb{E}_{z \sim \mathcal{D}}[p_z^{\leq N}]$.

4.4 Learning with and without quantum memory

We begin with the simpler task of learning to predict properties in an unknown quantum system ρ . We consider the setting where an algorithm has access to a physical source (an oracle) that generates an unknown quantum state ρ . The goal is to learn to predict properties of ρ from as

few accesses to the oracle as possible. We denote the number of oracle accesses to be T . We give definitions for learning algorithms that can utilize quantum memory and those that can not utilize quantum memory throughout the experiments.

Definition 4.15 (Learning states without quantum memory). *The algorithm can obtain classical data from the oracle by performing arbitrary POVM measurements on ρ . For each access to the oracle, the algorithm can select a POVM $\{F_s\}_s$ and obtain the classical outcome s with probability $\text{tr}(F_s\rho)$. The selection of the POVM can depend on all previous measurement outcomes. After T oracle accesses, the algorithm predicts the properties of ρ .*

Definition 4.16 (Learning states with quantum memory). *The algorithm can obtain quantum state ρ from the oracle and store in the quantum memory. After T oracle accesses, the algorithm performs a POVM $\{F_s\}_s$ on $\rho^{\otimes T}$ to predict properties of ρ .*

We now describe the general task of learning to predict properties in an unknown quantum channel \mathcal{C} . We consider the setting where an algorithm can access an unknown quantum-mechanical process, given by the quantum channel \mathcal{C} . The goal is to learn to predict properties of \mathcal{C} from minimum number of accesses to \mathcal{C} . Similarly, we denote the number of oracle accesses by T . We also give definitions for learning algorithms with and without quantum memory.

Definition 4.17 (Learning channels without quantum memory). *The algorithm can obtain classical data from the oracle by preparing an arbitrary initial state $|\psi\rangle\langle\psi|$, evolve under \mathcal{C} to yield the output state $(\mathcal{C} \otimes \mathcal{I})(|\psi\rangle\langle\psi|)$, and perform an arbitrary POVM measurement on the output state $(\mathcal{C} \otimes \mathcal{I})(|\psi\rangle\langle\psi|)$. For each access to the oracle, the algorithm can select a POVM $\{F_s\}_s$ and obtain the classical outcome s with probability $\text{tr}(F_s(\mathcal{C} \otimes \mathcal{I})(|\psi\rangle\langle\psi|))$. The selection of the initial state and the POVM can depend on all previous measurement outcomes. After T oracle accesses, the algorithm predicts the properties of \mathcal{C} .*

Definition 4.18 (Learning channels with quantum memory). *The algorithm can access the oracle \mathcal{C} as a quantum channel during the quantum computation. In particular, we consider a mixed state quantum computation. The resulting quantum memory after T oracle accesses can be written by*

$$\rho_T^{\mathcal{C}} = \mathcal{C}_T(\mathcal{C} \otimes \mathcal{I}) \dots \mathcal{C}_2(\mathcal{C} \otimes \mathcal{I}) \mathcal{C}_1(\mathcal{C} \otimes \mathcal{I})(\rho_0), \quad (24)$$

where ρ_0 is the initial state of the quantum memory with arbitrary number of qubits, \mathcal{C}_t is a quantum channel for all $t = 1, \dots, T$. After T oracle accesses, the algorithm performs a POVM $\{F_s\}_s$ on the quantum memory $\rho_T^{\mathcal{C}}$ to predict properties of \mathcal{C} .

Learning channels is more general than learning states. We can reduce to the definitions for learning states by considering quantum channels \mathcal{C} with 0-qubit input state and n -qubit output state. Furthermore, for learning states or channels with/without quantum memory, we can restrict the POVM to be rank-1 POVM because we never consider post-measurement quantum states in the definitions.

Remark 4.19. *In the definitions of learning states/channels with/without quantum memory, we can replace each POVM $\{F_s\}_s$ by a rank-1 POVM $\{\sqrt{w_s}2^n|\psi_s\rangle\langle\psi_s|\}_s$ using Lemma 4.8.*

Finally, we describe an intermediate regime where the algorithm has bounded quantum memory. We only consider definitions for learning states in the main text. The definition for learning channels with bounded quantum memory can be defined similarly. Intuitively, we consider the algorithms to have a k -qubit quantum memory σ . Every time the algorithm a new copy of quantum state

ρ , it performs a partial measurement on the combined state $\sigma \otimes \rho$ that results in a k -qubit post-measurement state. Because we need to consider post-measurement quantum state in this case, we *can not* replace the general POVM $\{F_s\}_s$ by a rank-1 POVM using Lemma 4.8. The formal definition is given below.

Definition 4.20 (Learning states with size- k quantum memory). *The algorithm maintains a k -qubit quantum memory σ . For each access to the oracle, the algorithm can select a POVM measurement $\{F_s\}_s$ on $(\rho \otimes \sigma)$, where each POVM element has some Cholesky decomposition $F_s = M_s^\dagger M_s$, and obtain a classical outcome s as well as the first k qubits of the post-measurement state,*

$$\text{tr}_{>k} \left(\frac{\mathcal{M}_s(\rho \otimes \sigma) \mathcal{M}_s^\dagger}{\text{tr}(F_s(\rho \otimes \sigma))} \right), \quad (25)$$

with probability $\text{tr}(F_s(\rho \otimes \sigma))$. For each access to the oracle, the algorithm can select a POVM $\{F_s\}_s$ that depends on all previous measurement outcomes. After T oracle accesses, the algorithm predicts the properties of ρ .

Remark 4.21. One can verify that the definition for learning states with quantum memory is equivalent to the definition for learning states with size- k quantum memory when $k \rightarrow \infty$.

4.5 Review of tensor network diagrams

It will be convenient to review the diagrammatic notation for tensor contraction, which we will leverage in several proofs. These so-called ‘tensor networks’ will render the index contraction of higher-rank tensors more transparent than standard notations. We also refer the interested reader to [Lan12, BC17] for a more comprehensive overview of tensor networks.

Diagrams for individual tensors

For our purposes, a rank (m, n) tensor is a multilinear map $T : \mathcal{H}^{*\otimes m} \otimes \mathcal{H}^{\otimes n} \rightarrow \mathbb{C}$. If $\{|i\rangle\}$ is an orthonormal basis for \mathcal{H} , then in bra-ket notation T can be expressed as

$$T = \sum_{\substack{i_1, \dots, i_m \\ j_1, \dots, j_n}} T_{j_1 \dots j_n}^{i_1 \dots i_m} (|i_1\rangle \otimes \dots \otimes |i_m\rangle) (\langle j_1| \otimes \dots \otimes \langle j_n|). \quad (26)$$

for some $T_{j_1 \dots j_n}^{i_1 \dots i_m} \in \mathbb{C}$. It is clear that a quantum state $|\Psi\rangle$ on \mathcal{H} is a rank $(1, 0)$ tensor, being a map from $\mathcal{H}^* \rightarrow \mathbb{C}$. Accordingly, its dual $\langle\Psi|$ is a $(0, 1)$ tensor. Moreover a matrix $M = \sum_{ij} M_j^i |i\rangle\langle j|$ is a $(1, 1)$ tensor. We elect to represent T diagrammatically by


(27)

which has m outgoing legs on the left and n incoming legs on the right. Each leg in the diagram may be associated with an index of the coefficients $T_{j_1 \dots j_n}^{i_1 \dots i_m}$. We set the convention that outgoing legs are ordered counter-clockwise and incoming legs are ordered clockwise. For instance, in (27) the top-left outgoing leg corresponds to i_1 , the leg below to i_2 , and so on. Likewise the top-right incoming leg corresponds to j_1 , the leg below to j_2 , and so on.

Tensor contraction

We next explain how to depict tensor network contractions diagrammatically. For sake of illustration, suppose we have a rank $(2, 1)$ tensor

$$A = \sum_{ijk} A_{jk}^i |i\rangle (\langle j| \otimes \langle k|) = \text{---} \boxed{A} \begin{matrix} \text{---} \\ \text{---} \end{matrix} \quad (28)$$

and a rank $(1, 2)$ tensor

$$B = \sum_{lmn} B_{\ell}^{mn} (|m\rangle \otimes |n\rangle) \langle \ell| = \begin{matrix} \text{---} \\ \text{---} \end{matrix} \boxed{B} \text{---} \quad (29)$$

Now suppose we want to compute the tensor network contraction corresponding to

$$\sum_{ijk} A_{jk}^i B_i^{jk}. \quad (30)$$

Here lower indices are contracted with upper indices because this represents contracting vectors with covectors. The contraction in (30) is depicted diagrammatically as

$$\text{---} \boxed{A} \begin{matrix} \text{---} \\ \text{---} \end{matrix} \boxed{B} \begin{matrix} \text{---} \\ \text{---} \end{matrix} \text{---} \quad (31)$$

Comparing the diagram with (30), we see that contracted indices corresponding to outgoing and incoming lines which are glued together. The fact that vectors are to be contracted with covectors is reflected in the fact that we are only allowed to glue together lines in a manner consistent with their orientations.

As another example, given a matrix $M = \sum_{ij} M_j^i |i\rangle \langle j|$, the trace can be written as

$$\text{tr}(M) = \sum_i M_i^i = \text{---} \boxed{M} \text{---} \quad (32)$$

If M_1, M_2, \dots, M_k are matrices, then the product $M_1 M_2 \cdots M_k$ is depicted by

$$\text{---} \boxed{M_1} \text{---} \boxed{M_2} \text{---} \cdots \text{---} \boxed{M_k} \text{---} \quad (33)$$

Multiplication by a scalar

Given a tensor T , multiplication by a scalar α is often denoted by αT . In our diagrammatic notation, we will simply write

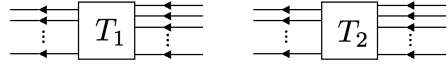
$$\alpha \begin{matrix} \text{---} \\ \vdots \\ \text{---} \end{matrix} \boxed{T} \begin{matrix} \text{---} \\ \vdots \\ \text{---} \end{matrix} \quad (34)$$

Tensor products

Given two tensors T_1, T_2 , we can form the tensor product $T_1 \otimes T_2$. We will denote this diagrammatically as

$$\begin{matrix} \text{---} \\ \vdots \\ \text{---} \end{matrix} \boxed{T_1} \begin{matrix} \text{---} \\ \vdots \\ \text{---} \end{matrix} \\ \text{---} \\ \vdots \\ \text{---} \end{matrix} \boxed{T_2} \begin{matrix} \text{---} \\ \vdots \\ \text{---} \end{matrix} \quad (35)$$

or also


(36)

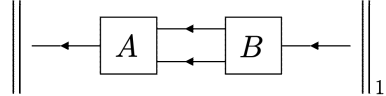
More generally, how to read off the order of a tensor product (e.g. $T_1 \otimes T_2$ or $T_2 \otimes T_1$) from a diagram will be clear in context.

Taking norms

Often it will be convenient to compute the norm of a matrix in tensor notation. For instance, if M is a matrix, then its 1-norm $\|M\|_1$ can be expressed diagrammatically as


(37)

Here we are simply taking the diagrammatic notation for M as a stand-in within the expression $\|M\|_1$. This is particularly convenient in circumstances where M is given by a tensor network contraction whose structure we wish to emphasize; for instance, the 1-norm of $M = \sum_{ijkl} A_{kl}^i B_j^{kl} |i\rangle\langle j|$ is conveniently depicted by


(38)

Tensors with legs of different dimensions

So far we have considered rank (m, n) tensors as maps $T : \mathcal{H}^{*\otimes m} \otimes \mathcal{H}^{\otimes n} \rightarrow \mathbb{C}$. More generally we can consider tensors $T : (\mathcal{H}_1^* \otimes \dots \otimes \mathcal{H}_m^*) \otimes (\mathcal{H}_{m+1} \otimes \dots \otimes \mathcal{H}_{m+n}) \rightarrow \mathbb{C}$ where the tensored Hilbert spaces in the domain need not be isomorphic. We can use the same diagrammatic notation as above, with the additional restriction that tensor legs can be contracted if they both carry the same dimension (i.e., correspond to a Hilbert space and a dual Hilbert space of the same dimension).


As an example, we can consider the state $|\Psi\rangle$ in $\mathbb{C}^2 \otimes \mathbb{C}^3$, and form its density matrix $|\Psi\rangle\langle\Psi|$. In our tensor diagram corresponding to this state, the \mathbb{C}^2 (qubit) legs will be solid lines and the \mathbb{C}^3 (qutrit) legs will be dotted lines. Performing a partial over the qutrit legs is expressed diagrammatically as


(39)

We will discuss the diagrammatic notation of partial traces in more detail below.

Identity operator

The identity operator on a Hilbert space \mathcal{H} can be expressed diagrammatically as an oriented line


(40)

We can clearly see that given a state in the Hilbert space


(41)

if we left-multiply by the identity diagram we will get the same tensor diagram and thus the same state. Likewise for the dual state


(42)

if we right-multiply by the identity diagram then we return the same tensor diagram.

Likewise, the identity operator on k copies of the Hilbert space $\mathcal{H}^{\otimes k}$ is just

$$\begin{array}{c} \text{---} \leftarrow \text{---} \\ \text{---} \leftarrow \text{---} \\ \vdots \\ \text{---} \leftarrow \text{---} \end{array} \quad (43)$$

In the setting that the Hilbert space under consideration is $\mathcal{H} \otimes \mathcal{H}'$ where each tensor factor has a different dimension, it is convenient to represent tensor legs in \mathcal{H} by solid lines and tensor legs in \mathcal{H}' by dotted line; in this setting the identity operator is

$$\begin{array}{c} \text{---} \leftarrow \text{---} \\ \cdots \cdots \leftarrow \cdots \cdots \end{array} \quad (44)$$

which readily generalizes if there are more than two Hilbert spaces with differing dimensions.

Resolutions of the identity

Suppose $\{|\Psi_i\rangle\}_i$ is an orthonormal basis for \mathcal{H} . Then the resolution of the identity $\sum_i |\Psi_i\rangle\langle\Psi_i| = \mathbb{1}$ can be expressed diagrammatically as

$$\sum_i \begin{array}{c} \leftarrow |\Psi_i\rangle\langle\Psi_i| \leftarrow \end{array} = \begin{array}{c} \text{---} \leftarrow \text{---} \end{array} \quad (45)$$

If instead $\{|\Psi_i\rangle\}_i$ is a resolution of the identity for $\mathcal{H} \otimes \mathcal{H}'$ where the two Hilbert spaces have different dimensions, we may analogously denote this diagrammatically by

$$\sum_i \begin{array}{c} \cdots \cdots \leftarrow |\Psi_i\rangle\langle\Psi_i| \leftarrow \cdots \cdots \end{array} = \begin{array}{c} \text{---} \leftarrow \text{---} \\ \cdots \cdots \leftarrow \cdots \cdots \end{array} \quad (46)$$

Similarly, if $\{M_s^\dagger M_s\}_s$ is a POVM on \mathcal{H} then the resolution of the identity $\sum_s M_s^\dagger M_s = \mathbb{1}$ can be written as

$$\sum_s \begin{array}{c} \leftarrow \boxed{\mathcal{M}_s^\dagger} \leftarrow \boxed{\mathcal{M}_s} \leftarrow \end{array} = \begin{array}{c} \text{---} \leftarrow \text{---} \end{array} \quad (47)$$

and analogously if the Hilbert space is $\mathcal{H} \otimes \mathcal{H}'$ or has even more tensor factors.

Taking traces and partial traces

Suppose we have a rank (n, n) tensor $T : \mathcal{H}^{*\otimes n} \otimes \mathcal{H}^{\otimes n}$. Then its trace is given by $\text{tr}(T) = \sum_{i_1, \dots, i_n} T_{i_1 \dots i_n}^{i_1 \dots i_n}$, or diagrammatically

$$\text{tr}(T) = \begin{array}{c} \text{---} \leftarrow \text{---} \\ \text{---} \leftarrow \text{---} \\ \vdots \\ \text{---} \leftarrow \text{---} \end{array} \quad (48)$$

A very useful diagrammatic identity is the trace of the identity matrix, which can be regarded as a rank $(1, 1)$ tensor $\mathbb{1} = \sum_i |i\rangle\langle i|$. We have

$$\text{tr}(\text{---} \leftarrow \text{---}) = \bigcirc = d \quad (49)$$

and so we see that a closed loop in tensor diagrams equals the dimension of the Hilbert space associated that curve. As another example, if we have the identity $\mathbb{1}_{d \times d} \otimes \mathbb{1}_{d' \times d'}$ on $\mathcal{H} \otimes \mathcal{H}'$, where

$\dim(\mathcal{H}) = d$, $\dim(\mathcal{H}') = d'$ and we have used subscripts on the identity matrices for emphasis, we have

$$\text{tr}(\text{---}\text{---}\text{---}) = \text{---}\text{---}\text{---} = d d' \quad (50)$$

where the solid line corresponds to the \mathcal{H} Hilbert space and the dotted line corresponds to the \mathcal{H}' Hilbert space.

We can also take partial traces in similar fashion. We define the partial trace over the ' k th subsystem' by

$$\text{tr}_k(T) = \sum_{\substack{i_1, \dots, i_{k-1}, i_{k+1}, \dots, i_n \\ j_1, \dots, j_{k-1}, j_{k+1}, \dots, j_n}} \left(\sum_{i_k} T_{j_1 \dots j_n}^{i_1 \dots i_n} \right) |i_1\rangle\langle j_1| \otimes \dots \otimes |i_{k-1}\rangle\langle j_{k-1}| \otimes |i_{k+1}\rangle\langle j_{k+1}| \otimes \dots \otimes |i_n\rangle\langle j_n|. \quad (51)$$

Note that $\text{tr}_\ell(\text{tr}_k(T)) = \text{tr}_k(\text{tr}_\ell(T))$. Since the operation of taking partial traces is commutative we can use the notation $\text{tr}_{k,\ell}(T)$. Notice that $\text{tr}_{1,\dots,n}(T) = \text{tr}(T)$. That is, taking the partial trace over all subsystems in the tensor is the same as taking the trace of the entire tensor.

Diagrammatically, the partial trace over the first subsystem is given by

$$\text{tr}_1(T) = \text{---}\text{---}\text{---} \quad (52)$$

The partial trace over the second subsystem is

$$\text{tr}_2(T) = \text{---}\text{---}\text{---} \quad (53)$$

and so on.

If we have an tensor with legs corresponding to Hilbert spaces of different dimensions, we can still in some cases take traces or partial traces. In particular, if $T : (\mathcal{H}_1^* \otimes \dots \otimes \mathcal{H}_n^*) \otimes (\mathcal{H}'_1 \otimes \dots \otimes \mathcal{H}'_m) \rightarrow \mathbb{C}$, then if $\mathcal{H}_k = \mathcal{H}'_k$ we can still compute the partial trace $\text{tr}_k(T)$. As a simple example consider the state $|\Psi\rangle$ living on $\mathcal{H} \otimes \mathcal{H}'$. Then its density matrix $|\Psi\rangle\langle\Psi|$ can be regarded as a $(2, 2)$ tensor taking $(\mathcal{H}^* \otimes \mathcal{H}'^*) \otimes (\mathcal{H} \otimes \mathcal{H}') \rightarrow \mathbb{C}$. Then we have

$$\text{tr}_2(|\Psi\rangle\langle\Psi|) = \text{---}\text{---}\text{---} \quad (54)$$

which is the same example as (39); a similar diagram expresses $\text{tr}_1(|\Psi\rangle\langle\Psi|)$.

Isotopies

We remark that tensor network diagrams are to be understood up to isotopy of the tensor legs; that is, deforming or bending the tensor legs does not change the interpretation of the diagram. For instance, for a product of matrices $M_1 M_2$ we have equivalences like

$$\text{---}\text{---}\text{---} = \text{---}\text{---}\text{---} \quad (55)$$

and similarly for all other kinds of tensors.

The isotopies are not required to be planar; for instance


(56)

We also can allow legs to cross, for instance


(57)

We will disregard whether such crossings are overcrossings or undercrossings.

However, we set the convention that we do not change the relative order of the endpoints of the outgoing or incoming legs. The reason is that permuting the order of the endpoints corresponds would correspond to permuting the tensor factors on which the tensor is defined. As a transparent example, let $T : (\mathcal{H}_1^* \otimes \mathcal{H}_2^*) \otimes (\mathcal{H}_1 \otimes \mathcal{H}_2) \rightarrow \mathbb{C}$ be denoted by


(58)

Then the diagram


(59)

corresponds to a tensor $(\mathcal{H}_2^* \otimes \mathcal{H}_1^*) \otimes (\mathcal{H}_1 \otimes \mathcal{H}_2) \rightarrow \mathbb{C}$ where we note that \mathcal{H}_1^* and \mathcal{H}_2^* have been permuted. See also the discussion of permutation operators below.

Permutation operators

Consider the permutation group on k elements, S_k , and let τ be an element of the group. We define a representation of τ , namely $\text{Perm}(\tau)$, which acts on a k -copy Hilbert space $\mathcal{H}^{\otimes k}$ as follows. Letting $|\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_n\rangle$ be a product state on $\mathcal{H}^{\otimes k}$, we define

$$\text{Perm}(\tau)|\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_n\rangle = |\psi_{\tau^{-1}(1)}\rangle \otimes |\psi_{\tau^{-1}(2)}\rangle \otimes \cdots \otimes |\psi_{\tau^{-1}(n)}\rangle \quad (60)$$

which extends to the entire Hilbert space $\mathcal{H}^{\otimes k}$ by linearity. With these conventions, the representations $\text{Perm}(\tau)$ enjoy the property

$$\text{Perm}(\tau) \cdot \text{Perm}(\sigma) = \text{Perm}(\tau\sigma) \quad (61)$$

where $\tau\sigma$ is shorthand for the group product, i.e. the composition $\tau \circ \sigma$.

These representations of S_k admit a very intuitive tensor diagrams. Consider, for instance, S_3 and $\tau = (123)$. Then the corresponding tensor diagram is


(62)

This is made very clear by labeling the endpoints of the diagram by


(63)

This notation generalized accordingly for other permutation representations. The group product structure is also transparent; for instance $\text{Perm}((123)) \cdot \text{Perm}((12))$ is depicted diagrammatically by


(64)

where $\text{Perm}((123))$ is given in red and $\text{Perm}((12))$ is given in blue for clarity; the allowed diagrammatic manipulations of performing isotopies without rearranging the endpoints of the tensor legs show that the result of the product is $\text{Perm}((23))$. A nice feature of the diagrams is that the diagram for $\text{Perm}(\tau^{-1})$ can be obtain from the diagram for $\text{Perm}(\tau)$ by flipping the latter horizontally.

As another example, if we multiply $\text{Perm}((123))$ by a state $|\Psi\rangle$ in $\mathcal{H}^{\otimes 3}$, then we get

$$\begin{array}{c} \text{---} \text{---} \text{---} \end{array} \Psi \rangle \quad (65)$$

from which it is clear that $\text{Perm}((123))$ permutes the tensor factors of the state according to $(123)^{-1} = (132)$.

In some later proofs where there is no ambiguity, we will denote $\text{Perm}(\tau)$ simply by τ .

Transposes and partial transposes

Suppose we have a matrix $M = \sum_{i,j} M_j^i |i\rangle\langle j|$ viewed as a rank $(1,1)$ tensor. We can represent its transpose $M^t = \sum_{i,j} M_j^i |j\rangle\langle i|$ diagrammatically by

Here we are dualizing each leg by changing the direction of each arrow, and then reorganizing the legs via isotopy so that the in-arrow comes in from the right and the out-arrow comes out to the left; this isotopy is done in order to match the arrow configuration in the diagram on the left.

$$\begin{array}{c} \text{---} \text{---} \end{array} M^t \begin{array}{c} \text{---} \text{---} \end{array} = \begin{array}{c} \text{---} \text{---} \end{array} M \begin{array}{c} \text{---} \text{---} \end{array} \quad (66)$$

If we have a higher-rank tensor, such as a rank $(2,2)$ tensor $T = \sum_{ijkl} T_{kl}^{ij} |i\rangle\langle k| \otimes |j\rangle\langle \ell|$, then we can also perform a partial transposition on a subsystem; for instance, the partial transposition on the second subsystem $\sum_{ijkl} T_{kl}^{ij} |i\rangle\langle k| \otimes |\ell\rangle\langle j|$ is given by

$$\begin{array}{c} \text{---} \text{---} \end{array} T \begin{array}{c} \text{---} \text{---} \end{array} \quad (67)$$

This notation extends to higher rank tensors in an analogous fashion.

Maximally entangled state

The maximally entangled state is given by $|\Omega\rangle = \sum_i |i\rangle|i\rangle$ where $\{|i\rangle\}$ is the computational basis. We treat $|\Omega\rangle$ as unnormalized, and it and its Hermitian conjugate are denoted by

$$\begin{aligned} |\Omega\rangle &= \begin{array}{c} \text{---} \text{---} \end{array} \bullet \\ \langle\Omega| &= \bullet \begin{array}{c} \text{---} \text{---} \end{array} \end{aligned} \quad (68)$$

Letting $\mathcal{H}_A \simeq \mathcal{H}_B \simeq \mathcal{H}_C$, we have the identities

$$(\mathbf{1}_A \otimes \langle\Omega|_{BC}) (|\Omega\rangle_{AB} \otimes \mathbf{1}_C) = \sum_i |i\rangle_A \langle i|_C \quad (69)$$

$$(\langle\Omega|_{AB} \otimes \mathbf{1}_C) (\mathbf{1}_A \otimes |\Omega\rangle_{BC}) = \sum_i |i\rangle_C \langle i|_A \quad (70)$$

which can be expressed diagrammatically as

$$\text{Diagram (71)} \quad (71)$$

$$\text{Diagram (72)} \quad (72)$$

We can think of the black dot as being a transpose operation since it changes the orientation of the tensor leg; moreover, two black dots annihilate one another since taking two transposes is the identity operation.

5 Exponential separations in learning quantum states

The main contribution of this work is in establishing information-theoretic lower bounds in learning algorithms without quantum memory. For learning properties of quantum states, a quantum-memoryless algorithm performs a POVM measurement on the unknown quantum state ρ and store the classical measurement outcome in a classical memory. The state of the classical memory is equivalent to the collection of all previous measurement outcomes. Because the POVM measurements performed at each time step depends on the previous outcomes, it is naturally to consider a tree representation for the learning algorithm. Each node on the tree represents the state of the classical memory. The classical memory state begins from the root of the tree. Every time we perform a POVM measurement on ρ , we move to a child node along the edge that encodes the measurement outcome. The POVM measurement we perform at each node will be different. Based on Remark 4.19, we only need to consider rank-1 POVM. Hence, we consider each node u to have an associated rank-1 POVM $\{\sqrt{w_s^u} 2^n |\psi_s^u\rangle\langle\psi_s^u|\}_s$. We also associate a probability to each node to represent the probability of arriving at the memory state in the learning process. Because we consider T measurements, all the leaf nodes are at depth T .

Definition 5.1 (Tree representation for learning states). *Fix an unknown n -qubit quantum state ρ . A learning algorithm without quantum memory can be expressed as a rooted tree \mathcal{T} of depth T , where each node on the tree encodes all the measurement outcomes the algorithm has seen so far. The tree satisfies the following properties.*

- Each node u is associated with a probability $p^\rho(u)$.
- For the root r of the tree, $p^\rho(r) = 1$.
- At each non-leaf node u , we measure a POVM $\{\mathcal{M}_s^u\}_s$ on ρ to obtain a classical outcome s . Each child node v of the node u is connected through the edge $e_{u,s}$.
- If v is the child node of u connected through the edge $e_{u,s}$, then

$$p^\rho(v) = p^\rho(u) w_s^u 2^n \langle \psi_s^u | \rho | \psi_s^u \rangle, \quad (73)$$

where $\{w_s^u 2^n |\psi_s^u\rangle\langle\psi_s^u|\}_s$ is a rank-1 POVM that depends on the node u .

- Every root-to-leaf path is of length T . Note that for a leaf node ℓ , $p^\rho(\ell)$ is the probability that the classical memory is in state ℓ after the learning procedure.

The general strategy we utilize is a reduction from the prediction task we care about to a two-hypothesis distinguishing problem. In particular, we consider distinguishing a set of unknown quantum state $\{\rho\}$ from the completely mixed state $\mathbb{1}/2^n$.

Definition 5.2 (Many-versus-one distinguishing task). *We consider the following two events to happen with equal probability.*

- The unknown state ρ is the n -qubit completely mixed state $\mathbb{1}/2^n$.
- The unknown state ρ is sampled uniformly from a set of n -qubit states $\{\rho_x\}_x$.

The goal of the learning algorithm is to predict which event has happened.

The basic strategy for proving lower bounds in a two-hypothesis distinguishing problem is the two-point method. After T measurements on the unknown quantum state ρ , the classical memory state will be represented by a leaf ℓ of the rooted tree \mathcal{T} . We will then use the information stored in the classical memory to distinguish between the two events. In order to successfully do so, intuitively, the distribution over the classical memory state in the two events must be sufficiently distinct. This is made rigorous by the famous two-point method in information-theoretic analysis. In the following lemma, we describe the two-point method in terms of the tree representation for a learning algorithm without quantum memory.

Lemma 5.3 (Le Cam's two-point method, see e.g. Lemma 1 in [Yu97]). *Consider a learning algorithm without quantum memory that is described by a rooted tree \mathcal{T} . The probability that the learning algorithm solves the many-versus-one distinguishing task correctly is upper bounded by*

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \left| \mathbb{E}_x p^{\rho_x}(\ell) - p^{\mathbb{1}/2^n}(\ell) \right|. \quad (74)$$

It is tempting to try to apply Lemma 5.3 by uniformly upper bounding the quantity $|\mathbb{E}_x p^{\rho_x}(\ell) - p^{\mathbb{1}/2^n}(\ell)|$ for all leaves ℓ . Unfortunately, it turns out that for some leaves, this quantity can be very large. The good news however is that we do have a uniform *one-sided* bound: as we will show, $p^{\mathbb{1}/2^n}(\ell) - \mathbb{E}_x p^{\rho_x}(\ell)$ (without the absolute values) can always be upper bounded by a very small value. It turns out that such a one-sided bound already suffices for applying Le Cam's two-point method.

Lemma 5.4 (One-sided bound suffices for Le Cam). *Consider a learning algorithm without quantum memory that is described by a rooted tree \mathcal{T} . If we have*

$$\frac{\mathbb{E}_x [p^{\rho_x}(\ell)]}{p^{\mathbb{1}/2^n}(\ell)} \geq 1 - \delta, \quad \forall \ell \in \text{leaf}(\mathcal{T}). \quad (75)$$

then the probability that the learning algorithm solves the many-versus-one distinguishing task correctly is upper bounded by δ .

Proof. We utilize the basic fact that $\frac{1}{2} \sum_i |p(i) - q(i)| = \sum_{i: p(i) \geq q(i)} p(i) - q(i)$, hence

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \left| \mathbb{E}_x p^{\rho_x}(\ell) - p^{\mathbb{1}/2^n}(\ell) \right| = \sum_{\substack{\ell \in \text{leaf}(\mathcal{T}) \\ p^{\mathbb{1}/2^n}(\ell) \geq \mathbb{E}_x p^{\rho_x}(\ell)}} p^{\mathbb{1}/2^n}(\ell) - \mathbb{E}_x p^{\rho_x}(\ell) \quad (76)$$

$$\leq \sum_{\substack{\ell \in \text{leaf}(\mathcal{T}) \\ p^{\mathbb{1}/2^n}(\ell) \geq \mathbb{E}_x p^{\rho_x}(\ell)}} p^{\mathbb{1}/2^n}(\ell) \delta \leq \delta. \quad (77)$$

We can then apply Lemma 5.3 and conclude the proof of the lemma. \square

5.1 Shadow tomography

Given M observables O_1, \dots, O_M with $\|O_i\|_\infty = 1, \forall i$, the goal of shadow tomography is to predict the expectation values $\text{tr}(O_i \rho)$ for all the observables up to ε -error. If we consider a class of traceless observables O_1, \dots, O_M such that

$$O_i = -O_{i+M/2}, \text{ and all eigenvalues of } O_i \text{ are } \pm 1, \quad \forall i = 1, \dots, M/2, \quad (78)$$

then the hardness of shadow tomography in learning algorithms without quantum memory can be characterized by the following quantity,

$$\delta(O_1, \dots, O_M) = \sup_{|\phi\rangle} \frac{2}{M} \sum_{i=1}^{M/2} \langle \phi | O_i | \phi \rangle^2, \quad (79)$$

where $\sup_{|\phi\rangle}$ is taken over the entire space of n -qubit pure states. We establish this hardness relation by the following theorem.

Theorem 5.5 (General shadow tomography lower bound). *Consider M traceless observables O_1, \dots, O_M satisfying Eq. (78), any learning algorithm without quantum memory requires*

$$T \geq \Omega \left(\frac{1}{\varepsilon^2 \delta(O_1, \dots, O_M)} \right) \quad (80)$$

copies of ρ to predict expectation values of $\text{tr}(O_x \rho)$ to at most ε -error for all $x = 1, \dots, M$ with at least a probability of $2/3$.

Proof. We consider a many-versus-one distinguishing task where we want to distinguish between maximally mixed state versus the set of n -qubit states $\{\rho_x\}_x$ given by

$$\rho_x = \frac{\mathbb{1} + 3\varepsilon O_x}{2^n}. \quad (81)$$

Using Eq. (78), we have $\text{tr}(O_x^2) = 2^n$. Using $\text{tr}(O_x^2) = 2^n$ and $\text{tr}(O_x) = 0$, we have $\text{tr}(O_x \rho_x) = 3\varepsilon$ and $\text{tr}(O_x (\mathbb{1}/2^n)) = 0$ for all $x = 1, \dots, M$. Therefore for any ρ_x , there exists an O_x with $\text{tr}(O_x \rho_x)$ substantially greater than zero. On the other hand, for $\mathbb{1}/2^n$, we have $\text{tr}(O_x (\mathbb{1}/2^n)) = 0$ for all O_x . Together, if a learning algorithm without quantum memory can predict the expectation values of $\text{tr}(O_x \rho), \forall x = 1, \dots, M$ up to error ε with probability at least $2/3$, then we can use the learning algorithm to perform the many-versus-one distinguishing task. Hence, a lower bound for this task also gives a lower bound for shadow tomography task. Given this result, we will now use Lemma 5.4 to establish a lower bound on T .

To utilize Lemma 5.4, we need to establish Eq. (75). Consider the tree representation \mathcal{T} of the learning algorithm without quantum memory given in Definition 5.1. For any leaf ℓ of \mathcal{T} , we consider the collection of edges on the path from root r to leaf ℓ to be $\{e_{u_t, s_t}\}_{t=1}^T$ where $u_1 = r$. From the recursive definition for the probability $p^\rho(u)$ in Definition 5.1, we have

$$p^\rho(\ell) = \prod_{t=1}^T w_{s_t}^{u_t} 2^n \langle \psi_{s_t}^{u_t} | \rho | \psi_{s_t}^{u_t} \rangle. \quad (82)$$

We then have the following calculation:

$$\frac{(\mathbb{E}_x p^{\rho_x}(\ell))}{p^{1/2^n}(\ell)} = \mathbb{E}_x \prod_{t=1}^T \left(\frac{w_s^u 2^n + 3\varepsilon w_s^u 2^n \langle \psi_{s_t}^{u_t} | O_x | \psi_{s_t}^{u_t} \rangle}{w_s^u 2^n} \right) \quad (83)$$

$$= \mathbb{E}_x \exp \left(\sum_{t=1}^T \log \left(1 + 3\varepsilon \langle \psi_{s_t}^{u_t} | O_x | \psi_{s_t}^{u_t} \rangle \right) \right) \quad (84)$$

$$\geq \exp \left(\sum_{t=1}^T \mathbb{E}_x \log \left(1 + 3\varepsilon \langle \psi_{s_t}^{u_t} | O_x | \psi_{s_t}^{u_t} \rangle \right) \right) \quad (85)$$

$$\geq \exp \left(\sum_{t=1}^T \frac{1}{M} \sum_{x=1}^{M/2} \log \left(1 - 9\varepsilon^2 \langle \psi_{s_t}^{u_t} | O_x | \psi_{s_t}^{u_t} \rangle^2 \right) \right) \quad (86)$$

$$\geq \exp \left(- \sum_{t=1}^T \frac{18}{M} \sum_{x=1}^{M/2} \varepsilon^2 \langle \psi_{s_t}^{u_t} | O_x | \psi_{s_t}^{u_t} \rangle^2 \right) \quad (87)$$

$$\geq \exp \left(- \sum_{t=1}^T 9\varepsilon^2 \delta(O_1, \dots, O_M) \right) \quad (88)$$

Equation (85) uses Jensen's inequality. Equation (86) uses the condition on the observables; see Equation (78). Equation (87) uses $\log(1-x) \geq -2x, \forall x \in [0, 0.79]$ which is satisfied given $\varepsilon < 0.29$. Equation (88) uses Equation (79). Plugging the above analysis to Eq. (87), we get

$$\frac{(\mathbb{E}_x p^{\rho_x}(\ell))}{p^{1/2^n}(\ell)} \geq \exp(-9T\varepsilon^2 \delta(O_1, \dots, O_M)) \geq 1 - 9T\varepsilon^2 \delta(O_1, \dots, O_M). \quad (89)$$

Utilizing Lemma 5.4 on a one-sided two-point method, we have that the learning algorithm without quantum memory solves the many-versus-one distinguishing task correctly with probability at most $9T\varepsilon^2 \delta(O_1, \dots, O_M)$. Combining with the fact that the learning algorithm without quantum memory solves the many-versus-one distinguishing task correctly with probability at least $2/3$, we have

$$T \geq \Omega \left(\frac{1}{\varepsilon^2 \delta(O_1, \dots, O_M)} \right). \quad (90)$$

This concludes the proof. \square

5.1.1 Lower bound for random observables

We consider a random collection of observables O_1, \dots, O_M on n -qubit systems given by

$$O_1 = U_1 Z_N U_1^\dagger, \dots, O_{M/2} = U_{M/2} Z_N U_{M/2}^\dagger, \quad (91)$$

$$O_{M/2+1} = -U_1 Z_N U_1^\dagger, \dots, O_M = -U_{M/2} Z_N U_{M/2}^\dagger, \quad (92)$$

where $U_1, \dots, U_{M/2}$ are sampled from the Haar measure over the unitary group $\mathbb{U}(2^n)$ and $Z_n = \mathbb{1} \otimes \dots \otimes Z$ is the Pauli-Z operator on the n -th qubit. It is easy to check that this collection of observables satisfies Eq. (78). This random collection of observables is very hard to predict.

Lemma 5.6 (δ for random observables). *With a constant probability over the random collection of the observations $\{O_1, \dots, O_M\}$, we have*

$$\delta(O_1, \dots, O_M) = \sup_{|\phi\rangle} \frac{2}{M} \sum_{i=1}^{M/2} \langle \phi | O_i | \phi \rangle^2 \leq \frac{2304\pi^3 \log(1+2M)}{M} + \frac{4}{M} + \frac{1}{2^n + 1}, \quad (93)$$

where $\sup_{|\phi\rangle}$ is taken over the entire space of n -qubit pure states.

Proof. We prove this lemma by the following strategy: take the covering net over all pure states, obtain exponential concentration inequality, and apply union bound. We first note that because $U_1, \dots, U_{M/2}$ are independent random matrices, for a fixed $|\phi\rangle$, we can consider $M/2$ independent and identically distributed pure states $|\psi_1\rangle, \dots, |\psi_{M/2}\rangle$ sampled from the Haar measure. Thus we can write

$$\frac{\varepsilon^2}{M/2} \sum_{i=1}^{M/2} \langle \psi | U_i Z_n U_i^\dagger | \psi \rangle^2 = \frac{\varepsilon^2}{M/2} \sum_{i=1}^{M/2} \langle \psi_i | Z_n | \psi_i \rangle^2. \quad (94)$$

Furthermore we have the following bound on the Lipschitz continuity of the function $f(|\psi\rangle) = \langle \psi | Z_n | \psi \rangle$,

$$f(|\psi\rangle) - f(|\xi\rangle) = \langle \psi | Z_n | \psi \rangle - \langle \xi | Z_n | \xi \rangle \quad (95)$$

$$\leq (\langle \psi | - \langle \xi |) Z_n (|\psi\rangle + |\xi\rangle) \quad (96)$$

$$\leq 2\| |\psi\rangle - |\xi\rangle \|_2. \quad (97)$$

Using Levy's lemma and $\mathbb{E}[f](|\psi\rangle) = 0$, we have the following concentration inequality

$$\Pr[|f(|\psi\rangle)| \geq t] \leq 2 \exp\left(-\frac{2^n t^2}{36\pi^3}\right). \quad (98)$$

This means $f(|\psi\rangle)$ is a sub-Gaussian random variable with variance proxy $18\pi^3/2^n$. A well-known fact in high-dimensional probability is that $X^2 - \mathbb{E}[X]^2$ is a sub-exponential random variable with variance proxy $16\sigma^2$ if X is a sub-Gaussian random variable with variance proxy σ^2 . Thus $f(|\psi_i\rangle)^2 - \mathbb{E}[f](|\psi_i\rangle)^2$ is a sub-exponential random variable with variance proxy $\lambda = 288\pi^3/2^n$. Using Bernstein's inequality, we have

$$\Pr\left[\left|\frac{1}{M/2} \sum_{i=1}^{M/2} f(|\psi_i\rangle)^2 - \mathbb{E}[f](|\psi_i\rangle)^2\right| \geq t\right] \leq 2 \exp\left(-\frac{M}{2} \min\left(\frac{t^2}{\lambda^2}, \frac{t}{\lambda}\right)\right). \quad (99)$$

Using the second moment of Haar-random vectors, we have $\mathbb{E}[f](|\psi_i\rangle)^2 = 1/(2^n + 1)$. An η -covering net \mathcal{N}_η for n -qubit pure state is upper bounded by $(1 + 2/\eta)^{2 \times 2^n}$. We can apply union bound to control the fluctuations in the covering net, which gives rise to the following inequality.

$$\Pr\left[\exists |\phi\rangle \in \mathcal{N}_\eta, \left|\frac{1}{M/2} \sum_{i=1}^{M/2} \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle^2 - \frac{1}{2^n + 1}\right| \geq t\right] \quad (100)$$

$$\leq 2(1 + 2/\eta)^{2 \times 2^n} \exp\left(-\frac{M}{2} \min\left(\frac{t^2}{\lambda^2}, \frac{t}{\lambda}\right)\right). \quad (101)$$

Now, we can extend from the covering net to the entire space of pure states by noting the following Lipschitz continuity property

$$\frac{1}{M/2} \sum_{i=1}^{M/2} \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle^2 - \frac{1}{M/2} \sum_{i=1}^{M/2} \langle \psi | U_i Z_n U_i^\dagger | \psi \rangle^2 \quad (102)$$

$$= \frac{1}{M/2} \sum_{i=1}^{M/2} (\langle \phi | U_i Z_n U_i^\dagger | \phi \rangle - \langle \psi | U_i Z_n U_i^\dagger | \psi \rangle) (\langle \phi | U_i Z_n U_i^\dagger | \phi \rangle + \langle \psi | U_i Z_n U_i^\dagger | \psi \rangle) \quad (103)$$

$$\leq \frac{1}{M/2} \sum_{i=1}^{M/2} 2 \left| \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle - \langle \psi | U_i Z_n U_i^\dagger | \psi \rangle \right| \leq 4 \|\phi - \psi\|_2. \quad (104)$$

This gives rise to the following concentration inequality for any pure states

$$\Pr \left[\exists |\phi\rangle, \left| \frac{1}{M/2} \sum_{i=1}^{M/2} \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle^2 - \frac{1}{2^n + 1} \right| \geq t + 4\eta \right] \quad (105)$$

$$\leq 2(1 + 2/\eta)^{2 \times 2^n} \exp \left(-\frac{M}{2} \min \left(\frac{t^2}{\lambda^2}, \frac{t}{\lambda} \right) \right). \quad (106)$$

We will assume that $M < 2^n$. When $M \geq 2^n$, we can start using classical shadow (an approach for shadow tomography using single-copy incoherent measurement that only requires $\log(M)$ copies), so a lower bound of $M/\log(M)$ would not hold in that regime. Under the assumption, we choose

$$\eta = \frac{1}{M}, t = \lambda \times \frac{8 \cdot 2^n \log(1 + 2M)}{M} = \frac{2304\pi^3 \log(1 + 2M)}{M}, \quad (107)$$

to yield the following probability bound

$$\Pr \left[\exists |\phi\rangle, \left| \frac{1}{M/2} \sum_{i=1}^{M/2} \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle^2 - \frac{1}{2^n + 1} \right| \geq t + 4\eta \right] \leq 2 \exp(-2). \quad (108)$$

Therefore with constant probability, we have

$$\sup_{|\phi\rangle} \frac{2}{M} \sum_{i=1}^{M/2} \langle \phi | U_i Z_n U_i^\dagger | \phi \rangle^2 \leq t + 4\eta + \frac{1}{2^n + 1} = \frac{2304\pi^3 \log(1 + 2M)}{M} + \frac{4}{M} + \frac{1}{2^n + 1}. \quad (109)$$

This concludes the proof of the lemma. \square

We can combine Lemma 5.6 that upper bounds $\delta(O_1, \dots, O_M)$ for the random collection of observables and Theorem 5.5 to obtain the following corollary. This corollary also implies the existence of a collection of observables O_1, \dots, O_M such that any learning algorithm without quantum memory requires $\Omega(\min(M/\log(M), 2^n)/\varepsilon^2)$ copies of ρ to perform shadow tomography.

Corollary 5.7 (Shadow tomography lower bound for random observables). *With a constant probability over the random collection of the observables $\{O_1, \dots, O_M\}$, any learning algorithm without quantum memory requires*

$$T \geq \Omega \left(\frac{\min(M/\log(M), 2^n)}{\varepsilon^2} \right) \quad (110)$$

copies of ρ to predict expectation values of $\text{tr}(O_x \rho)$ to at most ε -error for all $x = 1, \dots, M$ with at least a probability of $2/3$.

5.1.2 Lower bound for Pauli observables

Recall that we define an n -qubit Pauli observable to be the tensor product of n operators from the set $\{\mathbb{1}, X, Y, Z\}$; see Definition 4.9. To construct a collection of observables that are traceless and satisfies Equation (86), we define the following collection of $2(4^n - 1)$ Pauli observables with plus and minus signs,

$$P_1, \dots, P_{4^n-1} \in \{+\sigma_1 \otimes \dots \otimes \sigma_n, \sigma_i \in \{\mathbb{1}, X, Y, Z\}, \forall i\} \setminus \{\mathbb{1} \otimes \dots \otimes \mathbb{1}\}, \quad (111)$$

$$P_{4^n}, \dots, P_{2(4^n-1)} \in \{-\sigma_1 \otimes \dots \otimes \sigma_n, \sigma_i \in \{\mathbb{1}, X, Y, Z\}, \forall i\} \setminus \{-\mathbb{1} \otimes \dots \otimes \mathbb{1}\}, \quad (112)$$

This collection of observable yields an exponentially small $\delta(P_1, \dots, P_{2 \times (4^n-1)})$.

Lemma 5.8 (δ for Pauli observables).

$$\delta(P_1, \dots, P_{2 \times (4^n-1)}) = \sup_{|\phi\rangle} \frac{1}{4^n - 1} \sum_{i=1}^{4^n-1} \langle \phi | P_i | \phi \rangle^2 = \frac{1}{2^n + 1}. \quad (113)$$

Proof. For any n -qubit pure state $|\phi\rangle$, we have

$$\frac{1}{4^n - 1} \sum_{i=1}^{4^n-1} \langle \phi | P_i | \phi \rangle^2 = \frac{1}{4^n - 1} \text{tr} \left(\left(\sum_{i=1}^{4^n-1} P_i \otimes P_i \right) |\phi\rangle\langle\phi|^{\otimes 2} \right) \quad (114)$$

$$= \frac{1}{4^n - 1} \text{tr} \left((2^n \text{SWAP}_n - \mathbb{1}^{\otimes 2n}) |\phi\rangle\langle\phi|^{\otimes 2} \right) \quad (115)$$

$$= \frac{2^n - 1}{4^n - 1} = \frac{1}{2^n + 1}. \quad (116)$$

The second equality follows from Lemma 4.10. \square

We can combine Lemma 5.8 that characterizes $\delta(P_1, \dots, P_{2 \times (4^n-1)})$ for the collection of Pauli observables and Theorem 5.5 to obtain the following corollary.

Corollary 5.9 (Shadow tomography lower bound for Pauli observables). *Any learning algorithm without quantum memory requires*

$$T \geq \Omega(2^n / \varepsilon^2) \quad (117)$$

copies of ρ to predict expectation values of $\text{tr}(P_i \rho)$ to at most ε -error for all $i = 1, \dots, 2(4^n - 1)$ with at least a probability of $2/3$.

5.1.3 Upper bound

We give a upper bound for performing shadow tomography using a learning algorithm with only classical memory. There are two regimes. When $M \leq 2^n$, we apply the naive strategy of measuring each observable sequentially. When $M > 2^n$, we apply the classical shadow protocol [HKP20].

Theorem 5.10 (General shadow tomography upper bound). *For any collection of observables $\{O_1, \dots, O_M\}$ on n -qubit state with $\|O_i\|_\infty \leq 1$, there is a learning algorithm without quantum memory that uses*

$$T \leq \mathcal{O} \left(\frac{\min(M \log(M), 2^n \log(M))}{\varepsilon^2} \right) \quad (118)$$

copies of n -qubit state ρ to predict expectation values of $\text{tr}(O_x \rho)$ to at most ε -error for all $x = 1, \dots, M$ with high probability.

Proof. When $M \leq 2^n$, we measure each observable O_x on the unknown state ρ sequentially. For each observable O_x , we measure O_x on ρ for $\mathcal{O}(\log(M)/\varepsilon^2)$ times, such that the empirical average is close to $\text{tr}(O_x\rho)$ up to ε error with probability at least $1 - 0.01/M$. This gives a total of $T = \mathcal{O}(M \log(M)/\varepsilon)$ measurements on T copies of ρ . A union bound over all O_x shows that with probability at least 0.99, we can estimate all M expectation values $\text{tr}(O_1\rho), \dots, \text{tr}(O_M\rho)$ to ε error.

When $M > 2^n$, we apply the classical shadow protocol under random Clifford measurement given in [HKP20]. We perform a random Clifford measurement on each copy of ρ . Random Clifford measurement is equivalent to apply a random Clifford unitary sampled uniformly from the Clifford group to ρ , then measure in the computational basis. We record both the random Clifford unitary and the n -bit outcome from computational basis measurement. We perform the random Clifford measurement for $T = \mathcal{O}(2^n \log(M)/\varepsilon^2)$ times to gather a set of classical measurement data, also referred to as the classical shadow of ρ .

By combining Theorem 1 and Equation (S16) in [HKP20], we find that, for any collection of observables $\{O_1, \dots, O_M\}$ with $\text{tr}(O_x^2) \leq 2^n$, with high probability, we can use the classical measurement data (classical shadow of ρ) to estimate $\text{tr}(O_x\rho)$ to ε additive error for all $x = 1, \dots, M$. Because $\|O_i\|_\infty \leq 1$, we have $\text{tr}(O_x^2) \leq 2^n \|O_i\|_\infty^2 \leq 2^n$. Hence, we can apply classical shadow protocol with random Clifford measurement to achieve the stated complexity when $M > 2^n$. The stated result follows from combining the complexity from both $M \leq 2^n$ and $M > 2^n$. \square

We can see that the above upper bound matches with the lower bound for random observables stated in Corollary 5.7 up to poly-logarithmic scaling in M . When O_1, \dots, O_M are Pauli observables, we have $M = 2(4^n - 1)$ and the above upper bound becomes $\mathcal{O}(n2^n/\varepsilon^2)$, which matches the lower bound stated in Corollary 5.9 up to a factor of n .

5.2 Testing purity of the quantum state

5.2.1 Lower bound

In this subsection, we provide exponential lower bound for testing if a quantum state is pure or maximally mixed. In this section, we will denote $d = 2^n$ to be the Hilbert space dimension.

Theorem 5.11 (Purity testing lower bound). *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(2^{n/2}\right) \quad (119)$$

copies of $\rho \in \mathbb{H}^{2^n \times 2^n}$ to distinguish between whether ρ is a pure state or a maximally mixed state with probability at least $2/3$.

Proof. Let \mathcal{T} be the tree corresponding to any given learning algorithm for this distinguishing task. By Lemma 5.4 it suffices to lower bound $\mathbb{E}_v[p^{|\psi\rangle\langle\psi|}(\ell)/p^{\rho_{\text{mm}}}(\ell)]$ for all leaves ℓ . If $\{e_{u_t, s_t}\}_{t=1}^T$ are the edges on the path from root to the leaf ℓ , then

$$\mathbb{E}_v \left[\frac{p^{|\psi\rangle\langle\psi|}(\ell)}{p^{\rho_{\text{mm}}}(\ell)} \right] = \mathbb{E}_v \left[\prod_{t=1}^T d \langle \psi_{s_t}^{u_t} | \psi \rangle^2 \right] = \frac{d^T}{d(d+1) \cdots (d+T-1)} \cdot \sum_{\pi \in \mathcal{S}_T} \text{tr} \left(\pi \bigotimes_{t=1}^T |\psi_{s_t}^{u_t}\rangle\langle\psi_{s_t}^{u_t}| \right). \quad (120)$$

The second equality follows from Lemma 4.12 where π is the permutation operator. By Lemma 5.12 below, the sum in (120) is lower bounded by 1, so

$$\mathbb{E}_v \left[\frac{p^{|\psi\rangle\langle\psi|}(\ell)}{p^{\rho_{\text{mm}}}(\ell)} \right] \geq \frac{d^T}{d(d+1) \cdots (d+T-1)} \geq \prod_{t=0}^{T-1} \left(1 - \frac{t}{d}\right)^{-1} \geq \left(1 - \frac{T}{d}\right)^T. \quad (121)$$

5.2.2 Upper bound

Here we give a simple algorithm for the above distinguishing task that matches the lower bound in Theorem 5.11 up to constant factors.

Theorem 5.13. *There is a learning algorithm without quantum memory which takes $T = O(2^{n/2})$ copies of ρ to distinguish between whether ρ is a pure state or maximally mixed.*

To prove this, we will use the following well-known result from classical distribution testing:

Theorem 5.14 ([CDVV14, DKN14, CDGR18]). *Given $0 < \varepsilon < 1$ and sample access to a distribution q over $[d]$, there is an algorithm $\text{TESTUNIFORMITYL2}(q, d, \varepsilon)$ that uses $T = O(\sqrt{d}/\varepsilon^2)$ samples from q and with probability $9/10$ distinguishes whether q is the uniform distribution over $[d]$ or ε/\sqrt{d} -far in L_2 distance from the uniform distribution.*

We will also need the following standard moment calculation:

Lemma 5.15 (Lemma 6.4 in [CLO21]). *For Haar-random $\mathbf{U} \in U(2^n)$ and $\rho \in \mathbb{H}^{2^n \times 2^n}$, let Z denote the random variable $\sum_{i=1}^{2^n} (\langle i | \mathbf{U}^\dagger \mathbf{M} \mathbf{U} | i \rangle)^2$. Then*

$$\mathbb{E}[Z] = \frac{1}{2^n + 1} (\text{tr}(\mathbf{M})^2 + \|\mathbf{M}\|_{\text{HS}}^2). \quad (128)$$

If in addition we have that $\text{tr}(\mathbf{M}) = 0$, then

$$\mathbb{E}[Z^2] \leq \frac{1 + o(1)}{4^n} \|\mathbf{M}\|_{\text{HS}}^4. \quad (129)$$

We are now ready to prove Theorem 5.13.

Proof of Theorem 5.13. Sample a Haar-random basis $\{\mathbf{U} | i \rangle\}_{i \in [2^n]}$ and measure every copy of ρ in this basis. If ρ is maximally mixed, note that the distribution over outcomes from a single measurement is the uniform distribution u over $[2^n]$. On the other hand, if ρ is a pure state, let Z denote the random variable $\|q^{\mathbf{U}} - u\|_2^2$, where $q^{\mathbf{U}}$ is the distribution over outcomes from a single measurement. Note that Z is precisely the random variable Z defined in Lemma 5.15 for $\mathbf{M} = \rho - \rho_{\text{mm}}$, so we conclude that $\mathbb{E}[Z] = \frac{1}{2^n + 1} \cdot \|\rho - \rho_{\text{mm}}\|_{\text{HS}}^2$ and $\mathbb{E}[Z^2] \leq \frac{1 + o(1)}{16^n} \cdot \|\rho - \rho_{\text{mm}}\|_{\text{HS}}^4$, so by Paley-Zygmund, there is an absolute constant $c > 0$ for which $\Pr[|Z| \geq c \|\rho - \rho_{\text{mm}}\|_{\text{HS}}^2] \geq 9/10$. Note that $\|\rho - \rho_{\text{mm}}\|_{\text{HS}}^2 = 1 - 1/2^n$, so with probability at least $9/10$ over the randomness of \mathbf{U} , $\|q^{\mathbf{U}} - u\|_2 \geq \Omega(2^{-n/2})$.

So by Theorem 5.14, $\text{TESTUNIFORMITYL2}(q^{\mathbf{U}}, 2^n, \Theta(2^{-n/2}))$ will take $T = O(2^{n/2})$ samples from q and correctly distinguish whether q is uniform or far from uniform with probability at least $4/5$ over the randomness of the algorithm and of \mathbf{U} . \square

6 Exponential separation with bounded quantum memory

In this section, we consider a separation between algorithms with different quantum memory size. To the end, we will give a shadow tomography task and prove that any learning algorithm with k -qubit memory require $\Omega(2^{(n-k)/3})$ copies. In contrast, learning algorithms with n -qubit memory can succeed in the task using $\mathcal{O}(n)$ copies. In particular, the task is to predict the absolute value of the expectation $\text{tr}(P\rho)$ for all Pauli observables P in an unknown state ρ .

6.1 Lower bound with k -qubit of quantum memory

We restate the main theorem of this section, originally presented in Section 1.1.2, for convenience:

Theorem 1.4 (Shadow tomography with bounded quantum memory). *Any learning algorithm with k qubits of quantum memory requires $T \geq \Omega(2^{(n-k)/3})$ copies of ρ to predict $|\text{tr}(P\rho)|$ for all n -qubit Pauli observables P with at least probability $2/3$.*

To establish the theorem, we will again utilize the tree structure. The main difference is that each node on the tree is associated with a positive-semidefinite matrix that represents the quantum state of the k -qubit memory scaled by the probability of arriving at the node. In particular, we will work with the following analogue of Definition 5.1 for the bounded quantum memory setting:

Definition 6.1 (Tree representation for learning states with bounded quantum memory). *Fix an unknown n -qubit quantum state ρ . A learning algorithm with size- k quantum memory can be expressed as a rooted tree \mathcal{T} of depth T , where each node encodes the current state of the quantum memory in addition to the transcript of measurement outcomes the algorithm has seen so far. Specifically, the tree satisfies the following properties:*

1. *Each node u is associated with a k -qubit unnormalized mixed state $\Sigma^\rho(u)$ corresponding to the current state of the quantum memory.*
2. *For the root r of the tree, $\Sigma^\rho(r)$ is an initial state denoted by Σ_0 .*
3. *At each node u , we apply a POVM measurement $\{F_s^u\}_s$ on $\Sigma^\rho(u) \otimes \rho$ to obtain a classical outcome s . Each child node v of u is connected through the edge $e_{u,s}$.*
4. *For POVM element $F = M^\dagger M$ and $\Sigma \in \mathbb{H}^{2^k \times 2^k}$, define*

$$A_M^\rho(\Sigma) \triangleq \text{tr}_{>k} \left(M(\Sigma \otimes \rho) M^\dagger \right). \quad (130)$$

If v is the child node of u connected through the edge $e_{u,s}$, then

$$\Sigma^\rho(v) \triangleq A_{M_s^u}^\rho(\Sigma^\rho(u)). \quad (131)$$

$A_M^\rho(\Sigma)$ is the first k -qubit of the unnormalized post-measurement state from Definition 4.6.

5. *Note that for any node u , $p^\rho(u) \triangleq \text{tr}(\Sigma^\rho(u))$ is the probability that the transcript of measurement outcomes observed by the learning algorithm after t measurements is u . And $\Sigma^\rho(u)/p^\rho(u)$ is the state of the k -qubit memory at the node u .*

Given ρ , we will abuse notation and let p^ρ denote the distribution on leaves of \mathcal{T} given by the probabilities $\{p^\rho(\ell)\}_\ell$. Let P_1, \dots, P_{4^n-1} denote the collection of non-identity n -qubit Pauli observables from Definition 4.9; in this section, when we refer to a Pauli P we mean one from this collection, and we will use $\mathbb{E}_P[\cdot]$ to denote expectation with respect to a uniformly random such P .

We consider a many-versus-one distinguishing task where we want to distinguish between the completely mixed state ρ_{mm} versus the set of n -qubit states $\{\rho_P\}$ where P ranges over all n -qubit Pauli observables not equal to identity, where $\rho_P \triangleq (I + P)/2^n$. As in the proof of Theorem 5.5, a lower bound for this task immediately translates to one for shadow tomography. For the former, it suffices to show that for $T = o(2^{(n-k)/3})$, $d_{\text{TV}}(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}]) = o(1)$.

As in the proofs in Section 5, the primary technical ingredient for this will be a second moment bound. To formulate this, we will need the following object:

Definition 6.2. Given a POVM element $F = M^\dagger M$ and an unnormalized mixed state $\Sigma \in \mathbb{H}^{2^k \times 2^k}$, $\tau_{M,\Sigma} \in \mathbb{H}^{2^{n+k} \times 2^{n+k}}$ is given by

$$\tau_{M,\Sigma} = \text{Diagram} \quad . \quad (132)$$

(The dimensions of the Hilbert spaces corresponding to the edges have been labeled.)

One can think of the following lemma as bounding a matrix-valued analogue of the quantity defined in (113). Indeed, when $k = 0$, the following specializes up to constant factors to Lemma 5.8.

Lemma 6.3. For any POVM element M and unnormalized mixed state $\Sigma \in \mathbb{H}^{2^k \times 2^k}$,

$$\mathbb{E}_P[\|A_M^{\rho_{\text{mm}}}(\Sigma) - A_M^{\rho_P}(\Sigma)\|_{\text{tr}}^2] \leq \frac{1}{2^{n-k}} \cdot \frac{1}{2^{2n}-1} \cdot \text{tr}(\tau_{M,\Sigma})^2. \quad (133)$$

Proof. Note that $A_M^{\rho_{\text{mm}}}(\Sigma) - A_M^{\rho_P}(\Sigma) = -\text{tr}_{>k}(M(\frac{P}{2^n} \otimes \Sigma)M^\dagger)$, so by the fact that $\|X\|_{\text{tr}}^2 \leq 2^k \text{tr}(X^2), \forall X \in \mathbb{H}^{2^k \times 2^k}$, the left-hand side of (133) is upper bounded by

$$2^k \mathbb{E}_P \left[\text{tr} \left(\left[\text{tr}_{>k} \left(M \left(\frac{P}{2^n} \otimes \Sigma \right) M^\dagger \right) \right]^2 \right) \right]. \quad (134)$$

For fixed P , we express the expression inside the expectation diagrammatically as

$$\text{Diagram} = \text{Diagram} = \text{Diagram} \quad . \quad (135)$$

By Lemma 4.10, averaging (135) with respect to P yields

$$\frac{1}{2^n(2^{2n}-1)} \text{Diagram} = \frac{1}{2^{2n}(2^{2n}-1)} \text{Diagram} \leq \frac{1}{2^n(2^{2n}-1)} \text{Diagram} \quad (136)$$

where the inequality follows from the fact that the second term is equal to $\frac{1}{2^{2n}(2^{2n}-1)} \text{tr}(\text{tr}_{>k}(\tau)^2)$ which is non-negative. The claim follows from the fact that $\text{tr}(\tau^2) \leq \text{tr}(\tau)^2$ (as τ is positive-semidefinite) and utilizing Eqn. (134). \square

We will not be able to make use of the convexity trick present in the proof of Theorem 5.5 when each node is associated with a probability instead of positive-semidefinite matrix. Instead, we make use of a careful pruning argument; intuitively, for any leaf ℓ , we will essentially ignore the contribution to $d_{\text{TV}}(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}])$ coming from Paulis P for which $A_{M_s^u}^{\rho_P}$ behaves too differently to $A_{M_s^u}^{\rho_{\text{mm}}}$ for some edge $e_{u,s}$ on the path from root to ℓ .

Definition 6.4. A Pauli P is bad for an edge $e_{u,s}$ if

$$\left\| A_{M_s^u}^{\rho_{\text{mm}}}(\Sigma^{\rho_{\text{mm}}}(u)) - A_{M_s^u}^{\rho_P}(\Sigma^{\rho_{\text{mm}}}(u)) \right\|_{\text{tr}} \geq \frac{1}{2^{(n-k)/3}} \cdot \sqrt{\frac{1}{2^{2n}-1}} \cdot \text{tr}(\tau_{M_s^u, \Sigma^{\rho_{\text{mm}}}(u)}). \quad (137)$$

Otherwise we say P is good for $e_{u,s}$. Given node u , let $P[u]$ denote the set of all Paulis which are good for all edges on the path from root to u .

The following is an immediate consequence of Fact 6.3 and Markov's:

Fact 6.5. For any edge $e_{u,s}$, there are at most $2^{-(n-k)/3} \cdot (4^n - 1)$ bad Paulis $P \in \mathbb{H}^{2^n \times 2^n}$. In particular, along any given root-to-leaf path of the learning tree, there are at most $T \cdot 2^{-(n-k)/3} \cdot (4^n - 1)$ Paulis which are bad for some edge along the path.

Lemma 6.5 allows us to bound the $d_{\text{TV}}(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}])$ by a small term coming from bad Paulis and a term coming from good ones:

Lemma 6.6.

$$d_{\text{TV}}\left(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}]\right) \leq T \cdot 2^{-(n-k)/3} + \frac{1}{4^n - 1} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{P \in P[\ell]} \|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}} \quad (138)$$

Proof. Let \mathcal{L} denote the set of leaves ℓ for which $p^{\rho_{\text{mm}}}(\ell) \geq \mathbb{E}_P[p^{\rho_P}(\ell)]$. Then

$$d_{\text{TV}}\left(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}]\right) = \sum_{\ell \in \mathcal{L}} p^{\rho_{\text{mm}}}(\ell) - \mathbb{E}_P[p^{\rho_P}(\ell)] \quad (139)$$

$$\leq \sum_{\ell \in \mathcal{L}} \mathbb{E}_P[\min(p^{\rho_{\text{mm}}}(\ell), |p^{\rho_{\text{mm}}}(\ell) - p^{\rho_P}(\ell)|)] \quad (140)$$

$$\leq \sum_{\ell \in \mathcal{L}} \mathbb{E}_P[\min(p^{\rho_{\text{mm}}}(\ell), \|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}})] \quad (141)$$

$$\leq \sum_{\ell \in \mathcal{L}} \left[\Pr[P \notin P[\ell]] \cdot p^{\rho_{\text{mm}}}(\ell) + \frac{1}{4^n - 1} \sum_{P \in P[\ell]} \|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}} \right], \quad (142)$$

The first equality uses the fact that $d_{\text{TV}}(p, q) = \frac{1}{2} \sum_i |p_i - q_i| = \sum_{i: p_i \geq q_i} (p_i - q_i)$. Inequality (141) uses the fact that $\|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}} \geq \text{tr}(\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)) = p^{\rho_{\text{mm}}}(\ell) - p^{\rho_P}(\ell)$. The lemma follows from Fact 6.5 and the fact that $\sum_{\ell} p^{\rho_{\text{mm}}}(\ell) \leq 1$. \square

We are now ready to prove Theorem 1.4.

Proof of Theorem 1.4. By Lemma 6.6, it suffices to control the latter term on the right-hand side of Eq. (138). We do so via a hybrid argument. For any leaf ℓ with parent u and incoming edge $e_{u,s}$, and any $P \in P[\ell]$, we can upper bound $\|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}}$ by

$$\left\| A_{M_s^u}^{\rho_{\text{mm}}}(\Sigma^{\rho_{\text{mm}}}(u)) - A_{M_s^u}^{\rho_P}(\Sigma^{\rho_{\text{mm}}}(u)) \right\|_{\text{tr}} + \left\| A_{M_s^u}^{\rho_P}(\Sigma^{\rho_{\text{mm}}}(u)) - A_{M_s^u}^{\rho_P}(\Sigma^{\rho_P}(u)) \right\|_{\text{tr}}, \quad (143)$$

where we have used the transition formula in Eq. (131) and the triangle inequality. We can upper bound the former term by

$$\frac{1}{2^{(n-k)/3}} \cdot \sqrt{\frac{2^{2n}}{2^{2n}-1}} \cdot \text{tr} \left(\frac{1}{2^n} \tau_{M_s^u, \Sigma^{\rho_{\text{mm}}}(u)} \right) \quad (144)$$

because P is good for edge $e_{u,s}$; see Definition 6.4 and note that $\sqrt{2^{2n}} \cdot \frac{1}{2^n} = 1$. Using Definition 6.2 and the fact that $\{(M_s^u)^\dagger M_s^u\}_s$ is a POVM hence $\sum_s (M_s^u)^\dagger M_s^u = \mathbb{1}$, we have the following identity.

$$\sum_s \mathbb{E}_P \left[\text{tr} \left(\frac{1}{2^n} \tau_{M_s^u, \Sigma^{\rho_{\text{mm}}}(u)} \right) \right] = \mathbb{E}_P [\text{tr}(\Sigma^{\rho_{\text{mm}}}(u))], \quad \forall u : \text{node on } \mathcal{T}. \quad (145)$$

Therefore, we have

$$\sum_{u,s: \substack{e_{u,s} \text{ connected} \\ \text{to a leaf}}} \mathbb{E}_P \left[\text{tr} \left(\frac{1}{2^n} \tau_{M_s^u, \Sigma^{\rho_{\text{mm}}}(u)} \right) \right] = \mathbb{E}_P \left[\sum_{u: \substack{\text{connected} \\ \text{to a leaf}}} \text{tr}(\Sigma^{\rho_{\text{mm}}}(u)) \right] = 1. \quad (146)$$

As for the latter term in Eq. (143), note that for any leaf ℓ with parent u , $P[u] \subseteq P[\ell]$, so

$$\sum_{\ell, P \in P[\ell]} \left\| A_{M_s^u}^{\rho_P} (\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)) \right\|_{\text{tr}} \leq \sum_{u,s: \substack{e_{u,s} \text{ connected} \\ \text{to a leaf}, P \in P[u]}} \left\| A_{M_s^u}^{\rho_P} (\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)) \right\|_{\text{tr}} \quad (147)$$

$$= \sum_{u: \substack{u \text{ connected} \\ \text{to a leaf}, P \in P[u]}} \sum_s \left\| A_{M_s^u}^{\rho_P} (\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)) \right\|_{\text{tr}} \quad (148)$$

Note that for any node u , the map

$$\mathcal{E} : X \mapsto \sum_s |s\rangle\langle s| \otimes A_{M_s^u}^{\rho_P}(X) \quad (149)$$

is a quantum channel, so in particular $\|\mathcal{E}(X)\|_{\text{tr}} \leq \|X\|_{\text{tr}}$. In particular, we can see that

$$\|\mathcal{E}(\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u))\|_{\text{tr}} = \sum_s \left\| A_{M_s^u}^{\rho_P} (\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)) \right\|_{\text{tr}}. \quad (150)$$

We may thus upper bound Eq. (148) by

$$\sum_{u: \substack{\text{connected to} \\ \text{a leaf}, P \in P[u]}} \|\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)\|_{\text{tr}} = \sum_{u: \substack{\text{at depth } T-1 \\ P \in P[u]}} \|\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)\|_{\text{tr}}. \quad (151)$$

Combining Eq. (143), (144), and (151), we conclude that

$$\sum_{\substack{\ell \in \text{leaf}(\mathcal{T}) \\ P \in P[\ell]}} \|\Sigma^{\rho_{\text{mm}}}(\ell) - \Sigma^{\rho_P}(\ell)\|_{\text{tr}} \leq \frac{1}{2^{(n-k)/3}} \cdot \sqrt{\frac{2^{2n}}{2^{2n}-1}} + \sum_{u: \substack{\text{at depth } T-1 \\ P \in P[u]}} \|\Sigma^{\rho_{\text{mm}}}(u) - \Sigma^{\rho_P}(u)\|_{\text{tr}} \quad (152)$$

We can thus conclude by induction and by Lemma 6.6,

$$d_{\text{TV}} \left(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}] \right) \leq T \cdot 2^{-(n-k)/3} + T \cdot 2^{-(n-k)/3} \cdot \sqrt{\frac{2^{2n}}{2^{2n}-1}}. \quad (153)$$

In order to achieve the many-versus-one distinguishing task with probability at least $2/3$, we must have $2/3 \leq d_{\text{TV}}(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}])$ from Le Cam's two point method; see Lemma 5.3. However for $T = o(2^{(n-k)/3})$, $d_{\text{TV}}(p^{\rho_{\text{mm}}}, \mathbb{E}_P[p^{\rho_P}]) = o(1)$. This concludes the lower bound that $T \geq \Omega(2^{(n-k)/3})$. \square

6.2 Upper Bound with n -qubit of quantum memory

Here we show that with n qubits of quantum memory, we can estimate the absolute values of the expectations in M arbitrary Pauli observables using only $\mathcal{O}(\log M)$ copies of the unknown state.

Theorem 6.7. *Given unknown mixed state $\rho \in \mathbb{H}^{2^n \times 2^n}$ and any collection of M Pauli observables P_1, \dots, P_M , there is a learning algorithm using n qubits of quantum memory that measures $\mathcal{O}(\log(M)/\varepsilon^2)$ copies of ρ and outputs ε -accurate estimates for $|\text{tr}(P_i \rho)|^2$ for all $i \in [M]$ with probability at least $2/3$.*

Proof. We use an n -qubit memory to store a single copy of ρ from the previous experiment. In the current experiment, we can combine the quantum memory with the new quantum data ρ to measure the two copies $\rho \otimes \rho$ using an entangled measurement. For a pair of copies of ρ , if we measured every qubit j of $\rho \otimes \rho$ in the Bell basis to obtain outcome $|\beta_j\rangle \in \{|\Psi^+\rangle, |\Psi^-\rangle, |\Phi^+\rangle, |\Phi^-\rangle\}$, then note that because $\{P_i \otimes P_i\}_{i \in [M]}$ is simultaneously diagonalized in the n -fold tensor power of the Bell basis, we have

$$\mathbb{E}_{\{\beta_j\}} \left[\prod_{j=1}^n \text{tr}(P_{i(j)} \otimes P_{i(j)} |\beta_j\rangle\langle\beta_j|) \right] = \sum_{\{\beta_j\}} \prod_{j=1}^n \text{tr}(P_{i(j)} \otimes P_{i(j)} \Pi_{\beta_j} (\rho_j \otimes \rho_j) \Pi_{\beta_j}) = \text{tr}(P_i \rho)^2, \quad (154)$$

where $P_{i(j)}$ denotes the j -th qubit of P_i , ρ_j denotes the j -th qubit of ρ , and Π_{β} denotes the projector to Bell state β . The random variable in the expectation on the left-hand side of (154) is clearly bounded, so by Chernoff bounds, we can estimate its mean for any fixed j to within error ε with probability at least $1 - 1/3M$ using $\mathcal{O}(\log(M/\delta)/\varepsilon^2)$ samples. The claim follows by a union bound. \square

7 Exponential separations in learning quantum channels

7.1 Prerequisites

We begin by generalizing the tree representation for learning quantum states to the setting of learning quantum channels. First let us state the idea of the definition intuitively before delving into its technical description. There is some quantum channel \mathcal{C} which we wish to learn about; we have the ability to apply the channel to a state of our choice and then to completely measure the resulting state. The resulting measurement outcome can be recorded in a classical memory. The procedure of preparing a state, applying the channel, and then making a measurement is repeated over multiple rounds, wherein the measurement outcomes of previous rounds can inform the states prepared in future rounds, as well as the choice of measurement in future rounds. That is, the protocol is adaptive. At the end, we have gained a list of measurement outcomes with which we can judiciously infer properties of the channel \mathcal{C} under investigation.

Now we provide the full technical definition:

Definition 7.1 (Tree representation for learning channels). *Consider a fixed quantum channel \mathcal{C} acting on an n -qubit subsystem of a Hilbert space $\mathcal{H} \simeq \mathcal{H}_{\text{main}} \otimes \mathcal{H}_{\text{aux}}$ where $\mathcal{H}_{\text{main}} \simeq (\mathbb{C}^2)^{\otimes n}$ is the ‘main system’ comprising n qubits and $\mathcal{H}_{\text{aux}} \simeq (\mathbb{C}^2)^{\otimes n'}$ is an ‘auxiliary system’ of n' qubits. It is convenient to define $d = 2^n$ and $d' = 2^{n'}$. A learning algorithm without quantum memory can be represented as a rooted tree \mathcal{T} of depth T such that each node encodes all measurement outcomes the algorithm has received thus far. The tree has the following properties:*

- Each node u has an associated probability $p^{\mathcal{C}}(u)$.

- The root of the tree r has an associated probability $p^C(r) = 1$.
- At each non-leaf node u , we prepare a state $|\phi_u\rangle$ on \mathcal{H} , apply the channel \mathcal{C} onto the n -qubit subsystem, and measure a rank-1 POVM $\{\sqrt{w_s^u} dd' |\psi_s^u\rangle\langle\psi_s^u|\}_s$ (which can depend on u) on the entire system to obtain a classical outcome s . Each child node v of the node u corresponds to a particular POVM outcome s and is connected by the edge $e_{u,s}$. We refer to the set of child node of the node u as $\text{child}(u)$. Accordingly, we can relabel the POVM as $\{\sqrt{w_v} dd' |\psi_v\rangle\langle\psi_v|\}_{v \in \text{child}(u)}$.
- If v is a child node of u , then

$$p^C(v) = p^C(u) w_v dd' \langle\psi_v| (\mathcal{C} \otimes \mathcal{I}_{\text{aux}})[|\phi_u\rangle\langle\phi_u|] |\psi_v\rangle. \quad (155)$$

- Each root-to-leaf path is of length T . For a leaf of corresponding to node ℓ , $p^C(\ell)$ is the probability that the classical memory is in state ℓ after the learning procedure.

Each node u in the tree represents the state of the classical memory at one time step of the learning process. The associated probability $p^C(u)$ for a node u is the probability that the classical memory enters the state u during the learning process. Each time we perform one experiment, we transition from a node u to a child node of u .

There are several features of the definition which we will remark on. First and foremost, an important feature of the definition is that we have access to an auxiliary Hilbert space \mathcal{H}_{aux} for each state preparation and measurement. In particular, even though the channel \mathcal{C} only acts on n qubits, we can apply it to the first n qubits of a state $|\phi\rangle \in \mathcal{H}_{\text{main}} \otimes \mathcal{H}_{\text{aux}}$ which can be entangled between the n qubits and the auxiliary system. Moreover, we can measure the resulting state $(\mathcal{C} \otimes \mathcal{I}_{\text{aux}})[|\phi\rangle\langle\phi|]$ using POVM's which are entangled between the n qubits and the auxiliary system. The presence of the auxiliary system will render our proofs somewhat elaborate; moreover, the presence of the auxiliary system renders our results stronger than previous ones for adaptive incoherent access QUALMs [ACQ21]. In this particular QUALM setting, the notion of learning algorithm is similar, except that there is no auxiliary system.

We consider quantum channel learning tasks which were first studied in the QUALM setting without an auxiliary Hilbert space [ACQ21]. They are:

Definition 7.2 (Fixed unitary task). *Suppose that an n -qubit quantum channel \mathcal{C} is one of the following with equal probability:*

- \mathcal{C} is the completely depolarizing channel \mathcal{D} .
- \mathcal{C} is the unitary channel $\mathcal{C}[\rho] = U\rho U^\dagger$ for U a fixed, Haar-random unitary.

The fixed unitary task is to distinguish between the two above possibilities. We can also consider analogous versions of the problem where U is instead a Haar-random orthogonal matrix, or a Haar-random symplectic matrix.

Note that instead of considering the completely depolarizing channel \mathcal{D} can be thought of in a different way which makes the fixed unitary task more illuminating. Specifically, we can equivalently think of \mathcal{D} as an n -qubit unitary channel which applies an i.i.d. random Haar unitary each time the channel is applied. From this perspective, \mathcal{D} implements time-dependent random unitary dynamics (i.e. a new unitary is selected for each application of the channel); the task is then to distinguish this from time-independent random unitary dynamics wherein the channel applies a single fixed random unitary. Said more simply, from this point of view the task is to distinguish a type of time-dependent dynamics from a type of time-independent dynamics.

We also consider another task from [ACQ21] with a slightly different flavor:

Definition 7.3 (Symmetry distinction task). *Suppose that an n -qubit quantum channel \mathcal{C} is one of the following with equal probability:*

- $\mathcal{C}[\rho] = U\rho U^\dagger$ for U a fixed, Haar-random unitary matrix.
- $\mathcal{C}[\rho] = O\rho O^\dagger$ for O a fixed, Haar-random orthogonal matrix.
- $\mathcal{C}[\rho] = S\rho S^\dagger$ for S a fixed, Haar-random symplectic matrix.

The symmetry distinction task is to distinguish between the three above possibilities.

Unitary, orthogonal, and symplectic matrices manifest three different forms of what is called time-reversal symmetry [Dys62]. In this terminology, the symmetry distinction task is to determine the time-reversal symmetry class of \mathcal{C} . The task belongs to a class of problems of determining the symmetries of an uncharacterized system, which are important in experimental physics.

In the above distinguishing tasks, we will always reduce them to two-hypothesis distinguishing problem. We define a two-hypothesis distinguishing problem as follows.

Definition 7.4 (Two-hypothesis channel distinction task). *The following two events happen with equal probability:*

- The channel \mathcal{C} is sampled from a probability distribution D_A over channels.
- The channel \mathcal{C} is sampled from a probability distribution D_B over channels.

The goal is to distinguish whether \mathcal{C} is sampled from D_A or D_B .

For any two-hypothesis distinguishing problem, we can always apply the two-point method similar to Lemma 5.3 in learning quantum states.

Lemma 7.5 (Le Cam's two-point method, see e.g. Lemma 1 in [Yu97]). *Consider a learning algorithm without quantum memory that is described by a rooted tree \mathcal{T} . The probability that the learning algorithm solves the two-hypothesis channel distinction task correctly is upper bounded by*

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |(\mathbb{E}_{\mathcal{C} \sim D_A} p^{\mathcal{C}}(\ell)) - (\mathbb{E}_{\mathcal{C} \sim D_B} p^{\mathcal{C}}(\ell))|. \quad (156)$$

7.2 Review of the Weingarten calculus

Here we will review Haar measures on unitary, orthogonal, and symplectic matrices, and present several key lemmas that we will leverage in the later proofs. First we recall the definitions of orthogonal and symplectic matrices. Denoting the set of unitary matrices on $(\mathbb{C}^2)^{\otimes n}$ by

$$U(d) = \{U \in \text{Mat}_{d \times d}(\mathbb{C}) : U^\dagger = U^{-1}\}, \quad (157)$$

the set of orthogonal matrices on is given by

$$O(d) = \{O \in U(d) : O^t = O^{-1}\}, \quad (158)$$

and the set of symplectic matrices is given by

$$\text{Sp}(d/2) = \{S \in U(d) : JS^t J^{-1} = S^{-1}\}. \quad (159)$$

Here J is the symplectic form

$$J = \begin{bmatrix} \mathbf{0}_{d/2 \times d/2} & \mathbf{1}_{d/2 \times d/2} \\ -\mathbf{1}_{d/2 \times d/2} & \mathbf{0}_{d/2 \times d/2} \end{bmatrix}, \quad (160)$$

where $\mathbf{0}_{d/2 \times d/2}$ is the $d/2 \times d/2$ matrix of all zeroes and $\mathbf{1}_{d/2 \times d/2}$ is the $d/2 \times d/2$ identity matrix. The quantity $JS^t J^{-1}$ is sometimes called the ‘‘symplectic transpose’’ and denoted by S^D .

Note that $\text{Sp}(d/2)$ is sometimes called the ‘‘symplectic unitary group’’ to distinguish it from the group of symplectic matrices which need not be unitary. For the orthogonal group, the matrices O will necessarily be real. For the symplectic unitary group, the matrices S could be complex numbers. For our purposes, we will adopt standard terminology by dropping the word ‘unitary’ since the context is clear.

Since $U(d)$, $O(d)$, $\text{Sp}(d/2)$ are each compact Lie groups, they admit canonical Haar measures which are right and left-invariant under group multiplication. For instance, for $U(d)$, the Haar measure satisfies

$$\int_{U(d)} dU f(U) = \int_{U(d)} dU f(VU) = \int_{U(d)} dU f(UV) \quad (161)$$

for any $V \in U(d)$ and any $f(U)$. Analogous expressions hold for the Haar measures corresponding to $O(d)$ and $\text{Sp}(d/2)$. Such Haar integrals will be essential for our proofs, and so here we catalog important properties.

Now we turn to discussing more detailed properties of the Haar integrals. Our short overview will be based on [CM17, Gu13, Mat13, ACQ21]. Instead of using the integral notation, we will often use expectation values $\mathbb{E}_{U \sim \text{Haar}}[\cdot]$.

7.2.1 Haar averaging over $U(d)$

For our purposes it will be useful to study moments of the Haar ensemble, in particular

$$\mathbb{E}_{U \sim \text{Haar}} \left[U_{i_1 j_1} U_{i_2 j_2} \cdots U_{i_k j_k} \overline{U_{i'_1 j'_1} U_{i'_2 j'_2} \cdots U_{i'_k j'_k}} \right] = \sum_{\sigma, \tau \in S_k} \delta_{\sigma(I), I'} \delta_{\tau(J), J'} \text{Wg}^U(\sigma \tau^{-1}, d). \quad (162)$$

This equation requires some unpacking. On the left-hand side, the bar denotes complex conjugation; for instance $\overline{U_{ij}} = U_{ji}^\dagger$. On the right-hand side, S_k is the symmetric group on k elements, I is a multi-index $I = (i_1, \dots, i_k)$ and similarly for I', J, J' , and

$$\delta_{\sigma(I), I'} := \delta_{i_{\sigma(1)}, i'_1} \delta_{i_{\sigma(2)}, i'_2} \cdots \delta_{i_{\sigma(k)}, i'_k}. \quad (163)$$

Finally, $\text{Wg}^U(\cdot, d)$ is a map $S_k \rightarrow \mathbb{R}$ called the *unitary Weingarten function* to be specified shortly. To further compress notation, it will be convenient to fully commit to multi-index notation and write $U_{IJ}^{\otimes k} = U_{i_1 j_1} U_{i_2 j_2} \cdots U_{i_k j_k}$ so that (162) becomes

$$\mathbb{E}_{U \sim \text{Haar}} \left[U_{IJ}^{\otimes k} U_{KL}^{\dagger \otimes k} \right] = \sum_{\sigma, \tau \in S_k} \delta_{\sigma(I), I'} \delta_{\tau(J), J'} \text{Wg}^U(\sigma \tau^{-1}, d). \quad (164)$$

The utility of (164) is that it allows us to compute $\mathbb{E}_{\text{Haar}}[U^{\otimes k} \otimes U^{\dagger \otimes k}]$, and matrix elements thereof, in terms of data of the symmetric group on T elements. We remark that $\mathbb{E}_{\text{Haar}}[U^{\otimes k} \otimes U^{\dagger \otimes \ell}]$ vanishes for $k \neq \ell$, so (164) covers all non-trivial cases.

It still remains to specify the unitary Weingarten function $\text{Wg}^U(\cdot, d)$. In fact, it can be regarded as the inverse of an easily specified matrix. To this end, in a slight abuse of notation, we will let

permutations τ, σ in S_k also label their representations on $\mathcal{H}^{\otimes k}$. That is, τ will denote a unitary $d^k \times d^k$ matrix on $\mathcal{H}^{\otimes k}$ which permutes the k copies of \mathcal{H} according to the permutation specified by the label τ . Now we can readily define

$$G^U(\sigma\tau^{-1}, d) := \text{tr}(\sigma\tau^{-1}) = d^{\#(\sigma\tau^{-1})} \quad (165)$$

such that $\#(\sigma\tau^{-1})$ counts the number of cycles of $\sigma\tau^{-1}$. Now $\text{Wg}^U(\cdot, d)$ is defined by the identity

$$\sum_{\tau \in S^k} \text{Wg}^U(\sigma^{-1}\tau, d) G^U(\tau^{-1}\pi, d) = \delta_{\sigma, \pi} \quad (166)$$

for all $\sigma, \pi \in S_k$. This equation expresses that Wg^U and G^U are in fact inverses as $k! \times k!$ matrices. To see this more readily, we can use the notation $\text{Wg}^U(\sigma^{-1}\tau, d) = \text{Wg}_{\sigma, \tau}^U$ and $G^U(\tau^{-1}\pi, d) = G_{\tau, \pi}^U$ so that (166) is simply $\sum_{\tau} \text{Wg}_{\sigma, \tau}^U G_{\tau, \pi}^U = \delta_{\sigma, \pi}$.

We conclude by presenting a theorem, corollary, and lemma which will be used in our proofs.

Theorem 3.2 of [CM17]. *For any $\sigma \in S_k$ and $d > \sqrt{6}k^{7/4}$,*

$$\frac{1}{1 - \frac{k-1}{d^2}} \leq \frac{(-1)^{k-\#(\sigma)} d^{2k-\#(\sigma)} \text{Wg}^U(\sigma, d)}{\prod_i \frac{(2\ell_i-2)!}{(\ell_i-1)!\ell_i!}} \leq \frac{1}{1 - \frac{6k^{7/2}}{d^2}} \quad (167)$$

where the left-hand side inequality is valid for any $d \geq k$. Note that $\sigma \in S_k$ has cycle type (ℓ_1, ℓ_2, \dots) .

An immediate corollary is:

Corollary 7.6. $|\text{Wg}^U(\mathbb{1}, d) - d^{-k}| \leq \mathcal{O}(k^{7/2}d^{-(k+2)})$.

We also recapitulate a useful result from [ACQ21]:

Lemma 6 of [ACQ21]. $\sum_{\tau \in S_k} |\text{Wg}^U(\tau, d)| = \frac{(d-k)!}{d!}$.

7.2.2 Haar averaging over $O(d)$

Just as Haar averaging over $U(d)$ is intimately related to the permutation group S_k , Haar averaging over $O(d)$ (and likewise $\text{Sp}(d/2)$) is intimately related to pair partitions $P_2(2k)$. Accordingly, we will begin with discussing pair partitions.

Informally, a pair partition on $2k$ is a way of pairing off $2k$ elements (e.g., pairing off people in a ballroom dance class). There are in fact $\frac{(2k)!}{2^k k!} = (2k-1)!!$ possible pairings of $2k$ elements. More formally, a pair partition $\mathbf{m} \in P_2(2k)$ is a function $\mathbf{m} : [2k] \rightarrow [2k]$ satisfying $\mathbf{m}(2i-1) < \mathbf{m}(2i)$ for $1 \leq i \leq k$ and $\mathbf{m}(1) < \mathbf{m}(3) < \dots < \mathbf{m}(2k-1)$. The pair permutation is often notated as

$$\mathbf{m} = \{\mathbf{m}(1), \mathbf{m}(2)\} \{\mathbf{m}(3), \mathbf{m}(4)\} \dots \{\mathbf{m}(2k-1), \mathbf{m}(2k)\} \quad (168)$$

where the brackets denote individual pairs, and the constraints on \mathbf{m} order the pairs in a canonical (and unique) manner. In words, within each pair the ‘left’ element is always less than the ‘right’ element, and the pairs themselves are ordered according to the ‘left’ element of each pair.

It is natural to endow pair permutations with a group structure so that they form a subgroup M_{2k} of the permutation group S_{2k} . We simply define M_{2k} by

$$M_{2k} := \{\sigma \in S_{2k} : \sigma(2i-1) < \sigma(2i) \text{ for } 1 \leq i \leq k, \sigma(1) < \sigma(3) < \dots < \sigma(2k-1)\} \quad (169)$$

and it is readily checked that M_{2k} forms a group. We will often leverage the natural bijection between $P_2(2k)$, namely $\mathbf{m} \mapsto \sigma_{\mathbf{m}}$ where $\mathbf{m}(i) = \sigma_{\mathbf{m}}(i)$ for all i . The identity pairing is denoted by ϵ , and by the above bijection maps to the identity permutation $\sigma_{\epsilon} = \mathbb{1}$.

Pair permutations also have a notion of cycles, which differs from that of the symmetric group. That is, the cycle type of $\sigma_{\mathbf{m}}$ as an element of M_{2k} is different from the cycle type of $\sigma_{\mathbf{m}}$ thought of as an element of S_{2k} . To construct the pair partition cycles $\sigma_{\mathbf{m}}$ (we will also refer to this as the cycle type of \mathbf{m}), consider the function $f_{\mathbf{m}} : [2k] \rightarrow [2k]$ defined by

$$f_{\mathbf{m}}(i) = \begin{cases} \mathbf{m}(2j) & \text{if } i = \mathbf{m}(2j-1) \\ \mathbf{m}(2j-1) & \text{if } i = \mathbf{m}(2j) \end{cases}. \quad (170)$$

This function maps i to the integer it is paired with under \mathbf{m} . The function f_{ϵ} corresponds to the identity pairing. We can construct the cycles of \mathbf{m} as follows. Consider the sequence

$$(1, f_{\mathbf{m}}(1), f_{\epsilon} \circ f_{\mathbf{m}}(1), f_{\mathbf{m}} \circ f_{\epsilon} \circ f_{\mathbf{m}}(1), \dots). \quad (171)$$

This sequence is periodic, and so we truncate it at its period so that no element is repeated. We call this truncated list B_1 , and view it as a cyclically ordered list (i.e., the list is regarded as the same if it is cyclically permuted). If B_1 contains all of $[2k]$, then \mathbf{m} contains only one cycle, namely B_1 . Otherwise, let j be the smallest integer in $[2k]$ with is not in B_1 , and construct

$$(j, f_{\mathbf{m}}(j), f_{\epsilon} \circ f_{\mathbf{m}}(j), f_{\mathbf{m}} \circ f_{\epsilon} \circ f_{\mathbf{m}}(j), \dots). \quad (172)$$

This is likewise periodic, and we truncate it at its period to get the cyclically ordered list B_2 . If B_1 and B_2 do not contain all of $[2k]$, then we construct a B_3 , etc. When the procedure terminates, we have B_1, B_2, \dots which contain all of $[2k]$. The B_1, B_2, \dots are the pair partition cycles of \mathbf{m} . Their corresponding lengths b_1, b_2, \dots are all even, and the cycle type (also called the coset type) of \mathbf{m} is given by

$$(\mu_1, \mu_2, \dots) = (b_1/2, b_2/2, \dots) \quad (173)$$

which is often listed in descending order of cycle size.

Using our notation for pair partitions as well as the multi-index notation we established previously, we have the following formula for a Haar integral over the orthogonal group

$$\mathbb{E}_{O \sim \text{Haar}}[O_{IJ} O_{J'I'}^t] = \sum_{\mathbf{m}, \mathbf{n} \in P_2(2k)} \Delta_{\mathbf{m}}(II') \Delta_{\mathbf{n}}(JJ') \text{Wg}^O(\sigma_{\mathbf{m}}^{-1} \sigma_{\mathbf{n}}) \quad (174)$$

where II' merges the multi-indices I and I' as $II' = (i_1, i'_1, \dots, i_k, i'_k)$ and likewise for JJ' . Letting $II' = \mathbf{I} = (\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_{2k})$ we define

$$\Delta_{\mathbf{m}}(\mathbf{I}) := \prod_{s=1}^k \delta_{\mathbf{i}_{\mathbf{m}(2s-1)}, \mathbf{i}_{\mathbf{m}(2s)}}. \quad (175)$$

Similar to before, $\text{Wg}^O(\cdot, d)$ is a map $M_{2k} \rightarrow \mathbb{R}$ called the orthogonal Weingarten function. Although we have written (175) in a way that emulates (164), the formula (175) has a different character to it. Specifically, examining the left-hand side, we can equivalently write it as $\mathbb{E}_{\text{Haar}}[O_{II', JJ'}]$ where we have simply used the equivalence $O_{J'I'}^t = O_{I'J'}$. That is, (175) tells us how to compute $\mathbb{E}_{\text{Haar}}[O^{\otimes 2k}]$ and matrix elements thereof for $2k$ even; this integral vanishes if we replace $2k$ by an odd number. By contrast with the Haar unitary setting where we needed as many U 's as U^\dagger 's to get a non-trivial integral, here we just need an even number of O 's, essentially because $O^\dagger = O^t$.

Analogous to the unitary setting discussed above, we can define the orthogonal Weingarten function $\text{Wg}^O(\cdot, d)$ in terms of a simpler function

$$G^O(\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}}, d) := d^{\#^O(\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}})} \quad (176)$$

where $\#^O(\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}})$ counts the number of M_{2k} -cycles of $\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}}$. We have the identity

$$\sum_{\mathbf{n} \in P_2(2k)} \text{Wg}^O(\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}}, d) G^O(\sigma_{\mathbf{n}}^{-1}\sigma_{\mathbf{p}}, d) = \delta_{\mathbf{m}, \mathbf{p}}. \quad (177)$$

which expresses that Wg^O and G^O are inverses as $(2k-1)!! \times (2k-1)!!$ matrices.

Finally, we present a theorem, corollary, and lemma which we will leverage in our proofs about the symmetry distinction problem:

Theorem 4.11 of [CM17]. *For any $\sigma_{\mathbf{m}} \in M_{2k}$ and $d > 12k^{7/2}$,*

$$\frac{1 - \frac{24k^{7/2}}{d}}{1 - \frac{144k^7}{d^2}} \leq \frac{(-1)^{k-\#^O(\sigma_{\mathbf{m}})} d^{2k-\#^O(\sigma_{\mathbf{m}})} \text{Wg}^O(\sigma_{\mathbf{m}}, d)}{\prod_i \frac{(2\mu_i-2)!}{(\mu_i-1)!\mu_i!}} \leq \frac{1}{1 - \frac{144k^7}{d^2}} \quad (178)$$

where $\sigma_{\mathbf{m}}$ has M_{2k} -cycle type (μ_1, μ_2, \dots) .

We have the immediate corollary

Corollary 7.7. $|\text{Wg}^O(\sigma_{\mathbf{e}}, d) - d^{-k}| \leq \mathcal{O}(k^7 d^{-(k+2)})$.

Analogous to Lemma 6 of [ACQ21] written above, we have [ACQ21]:

Lemma 8 of [ACQ21]. $\sum_{\mathbf{m} \in P_2(2k)} |\text{Wg}^O(\sigma_{\mathbf{m}}, d)| = \frac{(d-2k)!!}{d!!}$.

7.2.3 Haar averaging over $\text{Sp}(d/2)$

Haar averaging over the $\text{Sp}(d/2)$ is very similar to the orthogonal setting. We have the identity

$$\mathbb{E}_{S \sim \text{Haar}} [S_{IJ} S_{J'I'}^t] = \sum_{\mathbf{m}, \mathbf{n} \in P_2(2k)} \Delta'_{\mathbf{m}}(II') \Delta'_{\mathbf{n}}(JJ') \text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}^{-1}\sigma_{\mathbf{n}}, d/2) \quad (179)$$

where

$$\Delta'_{\mathbf{m}}(\mathbf{I}) := \prod_{s=1}^k J_{\mathbf{i}_{\mathbf{m}(2s-1)}, \mathbf{i}_{\mathbf{m}(2s)}}. \quad (180)$$

Here J is the canonical symplectic form defined in (160). The symplectic Weingarten function $\text{Wg}^{\text{Sp}}(\cdot, d/2)$ taking $M_{2k} \rightarrow \mathbb{R}$ is a small modification of the orthogonal Weingarten function, namely

$$\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}, d/2) = (-1)^k \varepsilon(\sigma_{\mathbf{n}}) \text{Wg}^O(\sigma_{\mathbf{m}}, -d) \quad (181)$$

where $\varepsilon(\sigma_{\mathbf{m}})$ is the signature of $\sigma_{\mathbf{m}}$ thought of as an element of S_{2k} .

We give a theorem, corollary and lemma analogous to the ones for the orthogonal group above.

Theorem 4.10 of [CM17]. *For any $\sigma_{\mathbf{m}} \in M_{2k}$ and $d > 6k^{7/2}$, we have*

$$\frac{1}{1 - \frac{k-1}{(d/2)^2}} \leq \frac{d^{2k-\#^{\text{Sp}}(\sigma_{\mathbf{m}})} |\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}, d/2)|}{\prod_i \frac{(2\mu_i-2)!}{(\mu_i-1)!\mu_i!}} \leq \frac{1}{1 - \frac{6k^{7/2}}{(d/2)^2}} \quad (182)$$

where $\sigma_{\mathbf{m}}$ has M_{2k} -cycle type (μ_1, μ_2, \dots) and $\#^{\text{Sp}}(\sigma_{\mathbf{m}}) = \#^O(\sigma_{\mathbf{m}})$.

This has the direct corollary

Corollary 7.8. $|\text{Wg}^{\text{Sp}}(\sigma_{\epsilon}, d/2) - d^{-k}| \leq \mathcal{O}(k^{7/2} d^{-(k+2)})$.

From [ACQ21] we borrow the useful lemma:

Lemma 10 of [ACQ21]. $\sum_{\mathbf{m} \in P_2(2k)} |\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}, d/2)| = \prod_{j=0}^{k-1} \frac{1}{d+2j}$.

7.3 Depolarizing channel versus random unitary

7.3.1 Lower bound without quantum memory

We are now prepared to establish the following results:

Theorem 7.9 (Exponential hardness of fixed unitary task without quantum memory). *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(d^{1/3}\right), \quad (183)$$

to correctly distinguish between the completely depolarizing channel \mathcal{D} on n qubits from a fixed, Haar-random unitary channel $\mathcal{U}[\rho] = U\rho U^\dagger$ on n qubits with probability at least $2/3$.

Theorem 7.10 (Exponential hardness of fixed orthogonal matrix task without quantum memory). *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(d^{2/7}\right), \quad (184)$$

to correctly distinguish between the completely depolarizing channel \mathcal{D} on n qubits from a fixed, Haar-random orthogonal matrix channel $\mathcal{U}[\rho] = O\rho O^t$ on n qubits with probability at least $2/3$.

Theorem 7.11 (Exponential hardness of fixed symplectic matrix task without quantum memory). *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(d^{1/3}\right), \quad (185)$$

to correctly distinguish between the completely depolarizing channel \mathcal{D} on n qubits from a fixed, Haar-random symplectic matrix channel $\mathcal{U}[\rho] = S\rho S^D$ on n qubits with probability at least $2/3$.

Hence, we established an exponential lower bound when the algorithms do not have a quantum memory. The proofs of Theorems 7.9, 7.10, and 7.11 have many similarities; however, the first involves heavy use of the combinatorics of permutations, whereas the latter two involve heavy use of the combinatorics of pair permutations. As such, we will prove Theorems 7.9 first, followed by a simultaneous proof of Theorems 7.10 and 7.11.

We now turn to a proof of Theorems 7.9, 7.10, and 7.11 which are our main result about the fixed unitary problem. We note that a special case of these theorems were proved in [ACQ21], namely where the learning protocol (see Definition 7.1) does not have an auxiliary system, i.e. \mathcal{H}_{aux} is a trivial Hilbert space. The inclusion of an \mathcal{H}_{aux} of arbitrary size $d' = 2^{n'}$ is our main technical contribution here; our proof will follow the same contours as that of [ACQ21], but with substantive modifications and generalizations. Indeed, the original proof strategy leads to T bounds like (183), (184), and (185), but reduced by factors of d'^T , rendering the bounds useless even when $d' \sim \mathcal{O}(1)$ (but non-zero).

7.3.2 Proof of Theorem 7.9

Let us begin with a proof of Theorem 7.9:

Proof. The proof begins by utilizing Lemma 7.5, which gives

$$\frac{2}{3} \leq \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - (\mathbb{E}_{\mathcal{U}} p^{\mathcal{U}}(\ell))|. \quad (186)$$

The goal now would be to obtain an upper bound right hand side. It is convenient to establish some notation. Each root-to-leaf path in \mathcal{T} is specified by a sequence of vertices v_0, v_1, \dots, v_T where $v_0 = r$ is the root and $v_T = \ell$ is a leaf. Moreover, the leaf ℓ determines the entire root-to-leaf path: it is the shortest path from that leaf to the root. So knowing ℓ is the same as knowing the entire path $v_0 = r, v_1, \dots, v_{T-1}, v_T = \ell$. Recall from Eq. (155) and following the root-to-leaf path $v_0 = r, v_1, \dots, v_{T-1}, v_T = \ell$, the probability of the leaf ℓ under the channel \mathcal{C} is

$$p^{\mathcal{C}}(\ell) = \prod_{t=1}^T (w_{v_t} d d' \langle \psi_{v_t} | (\mathcal{C} \otimes \mathcal{I}_{\text{aux}}) [|\phi_{v_{t-1}}\rangle \langle \phi_{v_{t-1}}|] |\psi_{v_t}\rangle), \quad (187)$$

where each $|\phi_{v_{t-1}}\rangle, |\psi_{v_{t-1}}\rangle$ lives in (dd') -dimensional Hilbert space $\mathcal{H} \simeq \mathcal{H}_{\text{main}} \otimes \mathcal{H}_{\text{aux}}$ from Definition 7.1. To analyze the probability, we let

$$|\Phi_\ell\rangle = |\phi_{v_0}\rangle \otimes |\phi_{v_1}\rangle \otimes \cdots \otimes |\phi_{v_{T-1}}\rangle \quad (188)$$

$$|\Psi_\ell\rangle = |\psi_{v_1}\rangle \otimes |\psi_{v_2}\rangle \otimes \cdots \otimes |\psi_{v_T}\rangle \quad (189)$$

$$W_\ell = w_{v_1} w_{v_2} \cdots w_{v_T} \quad (190)$$

where $v_0 = r$ and $v_T = \ell$. Notice that $|\Phi_\ell\rangle$ and $|\Psi_\ell\rangle$ each live in $\mathcal{H}^{\otimes T}$. Recall from Definition 7.1, for any node u , the set $\{\sqrt{w_v dd'} |\psi_v\rangle\langle\psi_v|\}_{v \in \text{child}(u)}$ is a POVM. Hence, we have $\sum_{v \in \text{child}(u)} w_v dd' |\psi_v\rangle\langle\psi_v| = \mathbb{1}_{\text{main}} \otimes \mathbb{1}_{\text{aux}}$. It is not hard to use this fact to derive the following identity

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell |\Psi_\ell\rangle \langle \Psi_\ell| = (\mathbb{1}_{\text{main}} \otimes \mathbb{1}_{\text{aux}})^{\otimes T} \quad (191)$$

and so accordingly $\sum_{\ell \in \text{leaf}(\mathcal{T})} W_\ell = 1$.

With these notations at hand, we can write $p^{\mathcal{D}}(\ell)$ diagrammatically as

$$p^{\mathcal{D}}(\ell) = d'^T W_{\ell} \begin{array}{ccc} \langle \Phi_{\ell} | & \longrightarrow & | \Phi_{\ell} \rangle \\ & \curvearrowright & \\ \langle \Psi_{\ell} | & \longrightarrow & | \Psi_{\ell} \rangle \end{array} . \quad (192)$$

For $\mathbb{E}_{\text{Haar}}[p^{\mathcal{U}}(\ell)]$, we utilize the Haar averaging of unitary group discussed in Section 7.2.1 to obtain

$$\mathbb{E}_{\text{Haar}}[p^{\mathcal{U}}(\ell)] = (dd')^T \sum_{\sigma, \tau \in S_T} W_{\ell} \begin{array}{c} \langle \Phi_{\ell} | \longrightarrow \boxed{\sigma} \longrightarrow | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \longrightarrow \boxed{\tau^{-1}} \longrightarrow | \Psi_{\ell} \rangle \end{array} W_{\ell}^U(\tau \sigma^{-1}, d) \quad (193)$$

where the solid lines correspond to $\mathcal{H}_{\text{main}}^{\otimes T}$ and the dotted lines correspond to $\mathcal{H}_{\text{aux}}^{\otimes T}$. It is convenient to let $p_{\sigma, \tau}(\ell)$ denote the summand of (193). We now use the triangle inequality in combination

with the Cauchy-Schwarz inequality to write

$$\begin{aligned} & \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - \mathbb{E}_{\text{Haar}}[p^{\mathcal{U}}(\ell)]| \\ & \leq \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\mathbb{1}, \mathbb{1}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\sigma \neq \mathbb{1}} |p_{\sigma, \mathbb{1}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\tau \neq \mathbb{1}, \sigma} |p_{\sigma, \tau}(\ell)|. \end{aligned} \quad (194)$$

We will bound each term in turn.

First term

Applying Cauchy-Schwarz to the first term in (194) we find that it is less than or equal to

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| \left| \text{Wg}^U(\mathbb{1}, d) - \frac{1}{d^T} \right|. \quad (195)$$

We can remove the absolute values on the diagrammatic term since it is strictly positive; this enables us to perform the sum over leaves $\sum_{\ell \in \text{leaf}(\mathcal{T})}$ and via the identity given in (191). The first term in (194) can now be written as

$$\frac{d^T}{2} \left| \text{Wg}^U(\mathbb{1}, d) - \frac{1}{d^T} \right|. \quad (196)$$

Since by Corollary 7.6 we have $|\text{Wg}^U(\mathbb{1}, d) - \frac{1}{d^T}| \leq \mathcal{O}(T^{7/2}/d^{T+2})$ for $T < \left(\frac{d}{\sqrt{6}}\right)^{4/7}$, Eqn. (196) is $\mathcal{O}(T^{7/2}/d^2)$ and so

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\mathbb{1}, \mathbb{1}}(\ell)| \leq \mathcal{O}\left(\frac{T^{7/2}}{d^2}\right). \quad (197)$$

Second term

Next we treat the second term in (194), namely $\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\sigma \neq \mathbb{1}} |p_{\sigma, \mathbb{1}}(\ell)|$. Utilizing the Cauchy-Schwarz inequality, this term is less than or equal to

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\sigma \neq \mathbb{1}} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \boxed{\sigma} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| |\text{Wg}^U(\sigma^{-1}, d)|. \quad (198)$$

We can dissect the middle term in the summand using the Hölder inequality

$$\left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \boxed{\sigma} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| \leq \left\| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \square \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right\|_1 \|\sigma\|_{\infty} \quad (199)$$

and notice that $\|\sigma\|_{\infty} = 1$. Furthermore, using the fact that the matrix inside the 1-norm on the right-hand side is positive semi-definite, we can replace the 1-norm with the trace and find the

bound

$$\left\| \begin{array}{c} \langle \Phi_\ell | \rightarrow \boxed{} \rightarrow | \Phi_\ell \rangle \\ \langle \Psi_\ell | \rightarrow \boxed{} \rightarrow | \Psi_\ell \rangle \end{array} \right\|_1 = \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{} \rightarrow | \Psi_\ell \rangle} . \quad (200)$$

Thus (198) is upper bounded by

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\sigma \neq \mathbb{1}} (dd')^T W_\ell \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{} \rightarrow | \Psi_\ell \rangle} |\text{Wg}^U(\sigma^{-1}, d)| . \quad (201)$$

Then applying the identity (191) we are left with

$$\frac{d^T}{2} \sum_{\sigma \neq \mathbb{1}} |\text{Wg}^U(\sigma^{-1}, d)| \quad (202)$$

which is less than or equal to $\mathcal{O}(T^2/d)$ by Lemma 6 of [ACQ21]. In summary, we have

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\sigma \neq \mathbb{1}} |p_{\sigma, \mathbb{1}}(\ell)| \leq \mathcal{O}\left(\frac{T^2}{d}\right) . \quad (203)$$

Third term

Finally we treat the third term in (194), namely $\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\tau \neq \mathbb{1}, \sigma} |p_{\sigma, \tau}(\ell)|$, which is the most difficult case. Leveraging the Cauchy-Schwarz inequality, this term is upper bounded by

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\tau \neq \mathbb{1}, \sigma} (dd')^T W_\ell \left| \frac{\langle \Phi_\ell | \rightarrow \boxed{\sigma} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right| |\text{Wg}^U(\tau\sigma^{-1}, d)| . \quad (204)$$

Applying the Hölder inequality to the second term in the summand as before, we find

$$\left| \frac{\langle \Phi_\ell | \rightarrow \boxed{\sigma} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right| \leq \left\| \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right\|_1 \|\sigma\|_\infty \quad (205)$$

where $\|\sigma\|_\infty = 1$. We also use the convenient inequality

$$\left\| \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right\|_1 = \left\| \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow \boxed{\tau} \boxed{\tau^{-1}} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right\|_1 \leq \left\| \frac{\langle \Phi_\ell | \rightarrow \boxed{} \rightarrow \boxed{\tau^{-1}} \rightarrow | \Phi_\ell \rangle}{\langle \Psi_\ell | \rightarrow \boxed{\tau^{-1}} \rightarrow | \Psi_\ell \rangle} \right\|_1 \quad (206)$$

to reorganize the order of the tensor legs; we have used the Hölder inequality to go from the middle term to the last term, and the fact that $\|\tau\|_\infty = 1$.

For a fixed permutation τ^{-1} , we can decompose it into cycles $C_1 C_2 \cdots C_{\#(\tau^{-1})}$. We say that $i \rightarrow j$ is in the m th cycle C_m if $C_m = (\cdots i j \cdots)$. Using this notation, for fixed $\tau^{-1} = C_1 C_2 \cdots C_{\#(\tau^{-1})}$ and letting $v_0 = r, v_1, \dots, v_{T-1}, v_T = \ell$ be the root-to-leaf path terminating in ℓ , we have

Lemma 7.12. *We have the following identity represented using tensor network diagrams.*

$$\left\| \begin{array}{c} \langle \Phi_\ell | \text{---} \boxed{} \text{---} \boxed{\tau^{-1}} | \Phi_\ell \rangle \\ \langle \Psi_\ell | \text{---} \boxed{\tau^{-1}} \text{---} \boxed{} | \Psi_\ell \rangle \end{array} \right\|_1 = \prod_{m=1}^{\#(\tau^{-1})} \prod_{i \rightarrow j \in C_m} \left\| \begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{j-1}} \rangle \\ \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_j} \rangle \end{array} \right\|_1. \quad (207)$$

Proof. Take any cycle C_m and any $i \rightarrow j \in C_m$. The solid leg of $\langle \phi_{v_{i-1}} |$ is dangling and the dotted leg connects to $\langle \psi_{v_i} |$. Similarly, the solid leg of $|\phi_{v_{j-1}} \rangle$ is dangling and the dotted leg connects to $|\psi_{v_j} \rangle$. Lastly, the solid leg of $\langle \psi_{v_i} |$ connects to $|\psi_{v_j} \rangle$. This accounts for all legs among $\phi_{v_{i-1}}$, ψ_{v_i} , $\phi_{v_{j-1}}$, and ψ_{v_j} , so the part of the diagram on the left of (207) that corresponds to these four states is not connected to the rest of the diagram. In this fashion, we conclude that the diagram on the left is a tensor product of the diagrams on the left for all m and $i \rightarrow j \in C_m$, from which the lemma follows. \square

We would like to convert the product of trace norms in (207) into a product of traces. We do so via the following basic estimate. First, for any $i \in [T]$, define the unnormalized density operator $\tilde{\rho}_{v_i} \in \text{Mat}_{d \times d}(\mathbb{C})$ by

$$\tilde{\rho}_{v_i} := \begin{array}{c} \text{---} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{i-1}} \rangle \text{---} \\ \text{---} \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_i} \rangle \text{---} \end{array} \quad (208)$$

Lemma 7.13. *For any $i \rightarrow j \in C_m$, the corresponding term on the right-hand side of (207) is upper bounded by $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})$. In particular, (207) is at most*

$$\prod_{m=1}^{\#(\tau^{-1})} \prod_{i \rightarrow j \in C_m} \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})}. \quad (209)$$

Proof. Using the relations $\|A\|_1 = \|A \otimes A^\dagger\|_1^{1/2} \leq \|\text{SWAP} \cdot (A \otimes A^\dagger)\|_1^{1/2}$, we find

$$\left\| \begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{j-1}} \rangle \\ \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_j} \rangle \end{array} \right\|_1 \leq \left\| \begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{j-1}} \rangle \langle \phi_{v_{j-1}} | \text{---} \text{---} | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_j} \rangle \langle \psi_{v_j} | \text{---} \text{---} | \psi_{v_i} \rangle \end{array} \right\|_1^{1/2}.$$

Since the operator inside the 1-norm on the right-hand side is clearly positive semi-definite, we can replace the 1-norm with a trace, namely

$$\left(\begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{j-1}} \rangle \langle \phi_{v_{j-1}} | \text{---} \text{---} | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_j} \rangle \langle \psi_{v_j} | \text{---} \text{---} | \psi_{v_i} \rangle \end{array} \right)^{1/2}. \quad (210)$$

Observe that we can equivalently rewrite (210) as $\sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})}$, where $\tilde{\rho}_{v_i}$ is the component of the diagram denoted in red and $\tilde{\rho}_{v_j}$ is the one in blue, yielding the first part of the lemma. The second follows from plugging this into the right-hand side of Eqn. (207). \square

To bound (209), it is convenient to develop some notation for cycles C_m . The usual notation for a cycle of length p (which for us is less than or equal to T) is $C_m = (a_1 a_2 \cdots a_p)$ where in our setting $\{a_1, a_2, \dots, a_p\} \subseteq \{1, 2, \dots, T\}$. We will decorate each a_i by an additional subscript as $a_{m,i}$ to remember that it belongs to the m th cycle C_m . Similarly, we will sometimes write $p = |C_m|$ to remind ourselves that it depends on m . In this notation, we can write (209) as a product of

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,1}}} \tilde{\rho}_{v_{a_{m,2}}})} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,2}}} \tilde{\rho}_{v_{a_{m,3}}})} \cdots \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p-1}}} \tilde{\rho}_{v_{a_{m,p}}})} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}} \tilde{\rho}_{v_{a_{m,1}}})} \quad (211)$$

over the different $m = 1, \dots, \#(\tau^{-1})$. We can further process the above expression for a given m . We proceed by analyzing two cases: (i) $p = |C_m|$ is even, and (ii) $p = |C_m|$ is odd.

Case 1: p is even. Each term in Eqn. (211) has the form $\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}})}$ except for the last term; however, if we treat the i subscripts of $a_{m,i}$ modulo p , then we can write the last term as $\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}} \tilde{\rho}_{v_{a_{m,p+1}}})}$. We elect to use this notation. Then we can rearrange and group the terms in Eqn. (211) as follows

$$\left(\prod_{i \text{ odd}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}})} \right) \left(\prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}})} \right). \quad (212)$$

Using the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$, the above is upper bounded by

$$\frac{1}{2} \prod_{i \text{ odd}} \text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}) + \frac{1}{2} \prod_{j \text{ even}} \text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}}). \quad (213)$$

We will call the first term $\frac{1}{2} R_{m,-}$ and the second term $\frac{1}{2} R_{m,+}$.

Case 2: p is odd. Again consider the product in Eqn. (211). We can rearrange and group terms as

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}} \tilde{\rho}_{v_{a_{m,1}}})} \left(\prod_{\substack{i \text{ odd} \\ 1 \leq i \leq p-2}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}})} \right) \left(\prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}})} \right). \quad (214)$$

If two matrices A, B are Hermitian and positive semi-definite, then using Cauchy-Schwarz combined with operator norm inequalities we have $\text{tr}(AB) \leq \|A\|_2 \|B\|_2 \leq \|A\|_1 \|B\|_1 \leq \text{tr}(A) \text{tr}(B)$. Accordingly, we have $\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}} \tilde{\rho}_{v_{a_{m,1}}})} \leq \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}}) \text{tr}(\tilde{\rho}_{v_{a_{m,1}}})}$ and so (214) is upper bounded by

$$\left(\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}})} \prod_{\substack{i \text{ odd} \\ 1 \leq i \leq p-2}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}})} \right) \left(\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,1}}})} \prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}})} \right). \quad (215)$$

Again using the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$, we have the upper bound

$$\frac{1}{2} \text{tr}(\tilde{\rho}_{v_{a_{m,p}}}) \prod_{\substack{i \text{ odd} \\ 1 \leq i \leq p-2}} \text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}) + \frac{1}{2} \text{tr}(\tilde{\rho}_{v_{a_{m,1}}}) \prod_{j \text{ even}} \text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}}) \quad (216)$$

where we analogously call the first term $\frac{1}{2} R_{m,-}$ and the second term $\frac{1}{2} R_{m,+}$.

In both cases, note that $R_{m,-}$ and $R_{m,+}$ implicitly depend on the leaf ℓ , so we will denote them by $R_{m,-}^\ell$ and $R_{m,+}^\ell$ when we wish to make this dependence explicit.

Putting together our notation and bounds from Case 1 and Case 2, we find that Eqn. (209) is upper bounded by

$$\begin{aligned} \prod_{m=1}^{\#(\tau^{-1})} \prod_{i \rightarrow j \in C_m} \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})} &\leq \frac{1}{2^{\#(\tau^{-1})}} \prod_{m=1}^{\#(\tau^{-1})} (R_{m,-} + R_{m,+}) \\ &= \frac{1}{2^{\#(\tau^{-1})}} \sum_{i_1, \dots, i_{\#(\tau^{-1})} = \pm} R_{1,i_1} R_{2,i_2} \cdots R_{\#(\tau^{-1}), i_{\#(\tau^{-1})}}. \end{aligned} \quad (217)$$

Each term in the sum in the last line of (217) is a product of terms like $\text{tr}(\tilde{\rho}_{v_i})$ and $\text{tr}(\tilde{\rho}_{v_j} \tilde{\rho}_{v_k})$. But the key point is that we have arranged the equations so that each term in the sum has $\tilde{\rho}_{v_i}$ for each $i = 1, \dots, T$ appear *exactly once*. This has the following highly useful consequence.

Lemma 7.14. *Fix any $i_1, \dots, i_{\#(\tau^{-1})} \in \{+, -\}$. Then*

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell R_{1,i_1}^\ell R_{2,i_2}^\ell \cdots R_{\#(\tau^{-1}), i_{\#(\tau^{-1})}}^\ell \leq d^{\frac{T}{2} - \left\lfloor \frac{L(\tau^{-1})}{2} \right\rfloor} \quad (218)$$

where $L(\tau^{-1})$ is the length of the longest cycle in τ^{-1} .

Proof. For ease of notation, we will let $R_j^\ell \triangleq R_{j,i_j}^\ell$. Recall from (190) that $W_\ell = w_{v_1} w_{v_2} \cdots w_{v_T}$ and note that

$$\sum_{v: \text{depth}(v)=i} dd' w_v \tilde{\rho}_v = \langle \phi_{\text{parent}(v)} | \phi_{\text{parent}(v)} \rangle \mathbb{1}_{d \times d} = \mathbb{1}_{d \times d}. \quad (219)$$

Accordingly we have that for any $\rho \in \text{Mat}_{d \times d}(\mathbb{C})$,

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\tilde{\rho}_v \rho) = \text{tr}(\rho), \quad (220)$$

and in particular, for $\rho = \mathbb{1}$,

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\tilde{\rho}_v) = d. \quad (221)$$

We now turn to bounding the left-hand side of (218). Recalling that $\prod_j R_j^\ell$ as a product of terms like $\text{tr}(\tilde{\rho}_{v_i})$ and $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_{i'}})$, we will define some sets of indices encoding this data. Let $S_1^{(T)} \subseteq [T]$ denote the indices i for which $\text{tr}(\tilde{\rho}_{v_i})$ appears in $\prod_j R_j^\ell$, and let $S_2^{(T)} \subseteq [T] \times [T]$ denote the set of (unordered) pairs (i, i') for which $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_{i'}})$ appears, so that for any root-to-leaf path in \mathcal{T} consisting of nodes $v_1, \dots, v_T = \ell$, we have

$$\prod_j R_j^\ell = \prod_{i \in S_1^{(T)}} \text{tr}(\tilde{\rho}_{v_i}) \cdot \prod_{(i, i') \in S_2^{(T)}} \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_{i'}}) \quad (222)$$

by definition. Now construct $S_1^{(t)} \subseteq [t]$ and $S_2^{(t)} \subseteq [t] \times [t]$ for $1 \leq t < T$ inductively as follows. If $t \in S_1^{(t)}$ then define $S_1^{(t-1)} \triangleq S_1^{(t)} \setminus \{t\}$ and $S_2^{(t-1)} \triangleq S_2^{(t)}$. Otherwise if $(t, t') \in S_2^{(t)}$ for some $t' < t$, then define $S_1^{(t-1)} \triangleq S_1^{(t-1)} \cup \{t'\}$ and $S_2^{(t-1)} \triangleq S_2^{(t)} \setminus \{(t, t')\}$. We collect some basic observations about these two set sequences:

Observation 7.15. For every $i \in S_1^{(T)}$, we have that $i \in S_1^{(i)}$.

Observation 7.16. For every $(i, i') \in S_2^{(T)}$, if $i \leq i'$ then $i \in S_1^{(i)}$ while $i' \notin S_1^{(i')}$.

The reason for defining these set sequences is that we can extract $\tilde{\rho}_{v_T}$ from the product on the right-hand side of (222) and apply (221) (resp. (220)) if $T \in S_1^{(T)}$ (resp. $(T, t') \in S_2^{(T)}$ for some $t' < T$) to obtain

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell \prod_j R_j^\ell = d^{\mathbb{1}[T \in S_1^{(T)}]} \sum_{u: \text{depth}(u)=T-1} (dd')^{T-1} W_u \prod_{i \in S_1^{(T-1)}} \text{tr}(\tilde{\rho}_{v_i}) \cdot \prod_{(i, i') \in S_2^{(T-1)}} \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_{i'}}), \quad (223)$$

where $W_u = w_{v_1} \cdots w_{v_{T-1}}$ if the path from root to u in \mathcal{T} consists of $v_1, \dots, v_{T-1} = u$. Proceeding inductively, we can express the right-hand side of (223) as

$$d^{\sum_{t=1}^T \mathbb{1}[t \in S_1^{(t)}]}. \quad (224)$$

By Observations 7.15 and 7.16, $\sum_{t=1}^T \mathbb{1}[t \in S_1^{(t)}] = |S_1^{(T)}| + |S_2^{(T)}|$. Because every even cycle C_m of τ^{-1} contributes $|C_m|/2$ pairs to $S_2^{(T)}$, and every odd cycle C_m contributes $\lfloor |C_m|/2 \rfloor$ pairs to $S_2^{(T)}$ and one element to $S_1^{(T)}$, we conclude that

$$\begin{aligned} \sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell \prod_j R_j^\ell &= d^{\sum_{m=1}^{\#(\tau^{-1})} \lceil \frac{|C_m|}{2} \rceil} \\ &= d^{\frac{T}{2} - \sum_{m=1}^{\#(\tau^{-1})} \lfloor \frac{|C_m|}{2} \rfloor} \\ &\leq d^{\frac{T}{2} - \lfloor \frac{L(\tau^{-1})}{2} \rfloor} \end{aligned} \quad (225)$$

as claimed. \square

Putting all of our previous analysis together, in particular by combining Eqn.'s (205), (206), (207), (209), (217), and (218), we arrive at

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\tau \neq \mathbb{1}, \sigma} |p_{\sigma, \tau}(s)| \leq d^T \sum_{\sigma} |\text{Wg}^U(\sigma^{-1}, d)| \sum_{\tau \neq \mathbb{1}} d^{-\lfloor \frac{L(\tau^{-1})}{2} \rfloor}. \quad (226)$$

The first sum on the right-hand side is bounded by $d^T \sum_{\sigma} |\text{Wg}^U(\sigma^{-1}, d)| \leq 1 + \mathcal{O}\left(\frac{T^2}{d}\right)$. Considering the second sum on the right-hand side, let $N(T, \ell)$ be the number of permutations in S_T where the length of the longest cycle is ℓ . Then the second sum can be written as

$$\sum_{\ell=2}^T N(T, \ell) d^{-\lfloor \frac{\ell}{2} \rfloor} \quad (227)$$

where we omit $k = 1$ from the sum since it corresponds to the identity permutation. Since $N(T, \ell) \leq \binom{T}{\ell} \ell! = \frac{T!}{(T-\ell)!} < T^\ell$, (227) is upper bounded by

$$\sum_{\ell=2}^{\infty} T^\ell d^{-\lfloor \frac{\ell}{2} \rfloor} = \frac{(1+T) \frac{T^2}{d}}{1 - \frac{T^2}{d}} = \frac{T^3}{d} + \frac{T^2}{d} + \mathcal{O}\left(\frac{T^5}{d^2}\right). \quad (228)$$

Now if $T \leq o(d^{1/3})$, then this quantity is $o(1)$ for some absolute constant $c > 0$. Altogether, we find

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\tau \neq 1, \sigma} |p_{\sigma, \tau}(\ell)| \leq o(1). \quad (229)$$

□

7.3.3 Proof of Theorems 7.10 and 7.11

As discussed above, we will present proof of Theorems 7.10 and 7.11, making heavy use of pair partitions.

Proof. The probability distribution $p^{\mathcal{D}}(\ell)$ is notated the same way as before. We can depict $\mathbb{E}_{\text{Haar}}[p^{\mathcal{O}}(\ell)]$ diagrammatically by

$$\mathbb{E}_{\text{Haar}}[p^{\mathcal{O}}(\ell)] = (dd')^T \sum_{\mathbf{m}, \mathbf{n} \in P_2(2T)} W_{\ell} \begin{array}{c} \langle \Phi_{\ell} | \text{---} \boxed{\Delta_{\mathbf{m}}} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \boxed{\Delta_{\mathbf{n}}} \text{---} | \Psi_{\ell} \rangle \end{array} \text{Wg}^O(\sigma_{\mathbf{n}} \sigma_{\mathbf{m}}^{-1}, d) \quad (230)$$

where $p_{\mathbf{m}, \mathbf{n}}^{\mathcal{O}}(\ell)$ denotes the summand. Similarly $\mathbb{E}_{\text{Haar}}[p^{\mathcal{S}}(\ell)]$ is given by

$$\mathbb{E}_{\text{Haar}}[p^{\mathcal{S}}(\ell)] = (dd')^T \sum_{\mathbf{m}, \mathbf{n} \in P_2(2T)} W_{\ell} \begin{array}{c} \langle \Phi_{\ell} | \text{---} \boxed{\mathbf{J}^t} \boxed{\Delta'_{\mathbf{m}}} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \boxed{\mathbf{J}} \boxed{\Delta'_{\mathbf{n}}} \text{---} | \Psi_{\ell} \rangle \end{array} \text{Wg}^{\text{Sp}}(\sigma_{\mathbf{n}} \sigma_{\mathbf{m}}^{-1}, d/2) \quad (231)$$

where here $p_{\mathbf{m}, \mathbf{n}}^{\mathcal{S}}(\ell)$ denotes the summand. Moreover, $\mathbf{J} := J^{\otimes t}$. As before, we use the triangle inequality in combination with the Cauchy-Schwarz inequality to write the two inequalities

$$\begin{aligned} & \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - \mathbb{E}_{\text{Haar}}[p^{\mathcal{O}}(\ell)]| \\ & \leq \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\mathbf{e}, \mathbf{e}}^{\mathcal{O}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}} |p_{\mathbf{m}, \mathbf{e}}^{\mathcal{O}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}, \mathbf{n}} |p_{\mathbf{m}, \mathbf{n}}^{\mathcal{O}}(\ell)|. \end{aligned} \quad (232)$$

and

$$\begin{aligned} & \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - \mathbb{E}_{\text{Haar}}[p^{\mathcal{S}}(\ell)]| \\ & \leq \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\mathbf{e}, \mathbf{e}}^{\mathcal{S}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}} |p_{\mathbf{m}, \mathbf{e}}^{\mathcal{S}}(\ell)| + \frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}, \mathbf{n}} |p_{\mathbf{m}, \mathbf{n}}^{\mathcal{S}}(\ell)|. \end{aligned} \quad (233)$$

We will bound the right-hand sides of (232) and (233) term by term.

First term for $O(d)$ case

We apply the Cauchy-Schwarz inequality to the first term in (232) to find the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| \left| \text{Wg}^O(\sigma_{\mathbf{e}}, d) - \frac{1}{d^T} \right|. \quad (234)$$

As in the unitary setting, we remove the absolute values on the diagrammatic term by virtue of its positivity, and sum over leaves to get

$$\frac{d^T}{2} \left| \text{Wg}^O(\sigma_{\epsilon}, d) - \frac{1}{d^T} \right|. \quad (235)$$

Using Corollary 7.7 which gives us $|\text{Wg}^O(\sigma_{\epsilon}, d) - \frac{1}{d^T}| \leq \mathcal{O}(T^7/d^{T+2})$ for $T < (\frac{d}{12})^{2/7}$, we immediately find

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\epsilon, \epsilon}^{\mathcal{O}}(\ell)| \leq \mathcal{O}\left(\frac{T^7}{d^2}\right). \quad (236)$$

First term for $\text{Sp}(d/2)$ case

We recapitulate the same manipulations in the symplectic case. Applying the Cauchy-Schwarz inequality to the first term in (233) gives us the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| \left| \text{Wg}^{\text{Sp}}(\sigma_{\epsilon}, d/2) - \frac{1}{d^T} \right|. \quad (237)$$

We again remove the absolute values around the diagrammatic term, and sum over leaves to find

$$\frac{d^T}{2} \left| \text{Wg}^{\text{Sp}}(\sigma_{\epsilon}, d/2) - \frac{1}{d^T} \right|. \quad (238)$$

Using the analogous Corollary 7.8 which provides the bound $|\text{Wg}^O(\sigma_{\epsilon}, d) - \frac{1}{d^T}| \leq \mathcal{O}(T^{7/2}/d^{T+2})$ for $T < (\frac{d}{6})^{2/7}$, we have

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{D}}(\ell) - p_{\epsilon, \epsilon}^{\mathcal{S}}(\ell)| \leq \mathcal{O}\left(\frac{T^{7/2}}{d^2}\right). \quad (239)$$

Second term for $O(d)$ case

Applying Cauchy-Schwarz to the second term in (232), we find the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathfrak{m} \neq \epsilon} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \text{---} \boxed{\Delta_{\mathfrak{m}}} \text{---} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \text{---} \text{---} | \Psi_{\ell} \rangle \end{array} \right| |\text{Wg}^O(\sigma_{\mathfrak{m}}^{-1}, d)|. \quad (240)$$

Let us define the matrix

$$C(2i-1, 2i) := \begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_i} | \text{---} \text{---} | \psi_{v_i} \rangle \end{array} \quad (241)$$

where the $2i-1$ and $2i$ are just labels (i.e. they are not matrix indices). We will further define $C(2i, 2i-1) := C(2i-1, 2i)^t$. Then the diagrammatic term in (240) is equivalent to

$$\left| \text{tr} \left(\Delta_{\mathfrak{m}} \bigotimes_{i=1}^T C(2i-1, 2i) \right) \right|. \quad (242)$$

This can be expressed more explicitly as the absolute value of a product of traces of the C 's, namely

$$\left| \text{tr}(C(2, 1)C(f_{\mathbf{m}}(1), f_{\mathbf{e}} \circ f_{\mathbf{m}}(1))C(f_{\mathbf{m}} \circ f_{\mathbf{e}} \circ f_{\mathbf{m}}(1), f_{\mathbf{e}} \circ f_{\mathbf{m}} \circ f_{\mathbf{e}} \circ f_{\mathbf{m}}(1)) \cdots C(f_{\mathbf{e}} \circ f_{\mathbf{m}^{-1}}(2), f_{\mathbf{m}^{-1}}(2))) \right. \\ \left. \cdot \text{tr}(\cdots) \cdots \text{tr}(\cdots) \right| \quad (243)$$

where we have used the definition of $f_{\mathbf{m}}$ and $f_{\mathbf{e}}$ as per (170). Each trace corresponds to a particular M_{2T} -cycle of \mathbf{m} . Using the 1-norm inequality $\|A_1 A_2 \cdots A_k\|_1 \leq \prod_{i=1}^k \|A_i\|_1$, Eqn. (243) is upper bounded by

$$\prod_{i=1}^T \|C(2i-1, 2i)\|_1 \quad (244)$$

where we have used $\|C(2i-1, 2i)\|_1 = \|C(2i, 2i-1)\|_1$. Since each $C(2i-1, 2i)$ is positive semi-definite, $\|C(2i-1, 2i)\|_1 = \text{tr}(C(2i-1, 2i))$ and so (240) has the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}} (dd')^T W_{\ell} \begin{array}{c} \langle \Phi_{\ell} | \xrightarrow{\quad} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \xrightarrow{\quad} | \Psi_{\ell} \rangle \end{array} |\text{Wg}^O(\sigma_{\mathbf{m}}^{-1}, d)|. \quad (245)$$

Summing over leaves we arrive at

$$\frac{d^T}{2} \sum_{\mathbf{m} \neq \mathbf{e}} |\text{Wg}^O(\sigma_{\mathbf{m}}^{-1}, d)| \quad (246)$$

which is upper bounded by $\mathcal{O}(T^7/d^2) + \mathcal{O}(T^2/d)$ using Corollary 7.7 in combination with Lemma 8 of [ACQ21]. The ultimate result is

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{1}} |p_{\mathbf{m}, \mathbf{e}}^O(\ell)| \leq \mathcal{O}\left(\frac{T^7}{d^2}\right) + \mathcal{O}\left(\frac{T^2}{d}\right). \quad (247)$$

Second term for $\text{Sp}(d/2)$ case

A similar proof holds in the symplectic setting. We likewise apply Cauchy-Schwarz to the second term in (233) to obtain the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{e}} (dd')^T W_{\ell} \left| \begin{array}{c} \langle \Phi_{\ell} | \xrightarrow{\quad} \boxed{\mathbf{J}^t} \boxed{\Delta'_{\mathbf{m}}} | \Phi_{\ell} \rangle \\ \langle \Psi_{\ell} | \xrightarrow{\quad} | \Psi_{\ell} \rangle \end{array} \right| |\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}^{-1}, d/2)|. \quad (248)$$

Using the same notation for $C(2i-1, 2i)$ as in Eqn. (243) above, we further define

$$\tilde{C}(2i-1, 2i) := C(2i-1, 2i) \cdot J^t \quad (249)$$

and similarly $\tilde{C}(2i, 2i-1) := \tilde{C}(2i-1, 2i)^t$. Then the diagrammatic term in Eqn. (248) can be written as

$$\left| \text{tr} \left(\Delta'_{\mathbf{m}} \bigotimes_{i=1}^T \tilde{C}(2i-1, 2i) \right) \right|. \quad (250)$$

This can be expanded analogously to (243), namely as

$$\left| \text{tr}(\tilde{C}(2, 1) J C(f_m(1), f_\epsilon \circ f_m(1)) J \cdots J \tilde{C}(f_\epsilon \circ f_{m-1}(2), f_{m-1}(2)) J) \text{tr}(\cdots) \cdots \text{tr}(\cdots) \right|. \quad (251)$$

Using the same 1-norm inequality as the orthogonal case, Eqn. (251) is upper bounded by

$$\begin{aligned} \prod_{i=1}^T \|\tilde{C}(2i-1, 2i) J\|_1 &\leq \prod_{i=1}^T \|C(2i-1, 2i)\|_1 \|J\|_\infty \|J^t\|_\infty \\ &= \prod_{i=1}^T \text{tr}(C(2i-1, 2i)). \end{aligned} \quad (252)$$

In the first line we have used the Hölder inequality, and in the second line we used $\|J\|_\infty = \|J^t\|_\infty = 1$ as well as $\|C(2i-1, 2i)\|_1 = \text{tr}(C(2i-1, 2i))$. Thus we arrive at an upper bound for (248):

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{m \neq e} (dd')^T W_\ell \begin{array}{c} \langle \Phi_\ell | \longrightarrow | \Phi_\ell \rangle \\ \curvearrowright \\ \langle \Psi_\ell | \longleftarrow | \Psi_\ell \rangle \end{array} |\text{Wg}^{\text{Sp}}(\sigma_m^{-1}, d/2)|. \quad (253)$$

As before we sum over leaves, giving us

$$\frac{d^T}{2} \sum_{\mathbf{m} \neq \mathbf{e}} |\mathbf{W}_{\mathbf{g}}^{\text{Sp}}(\sigma_{\mathbf{m}}^{-1}, d/2)|. \quad (254)$$

This is upper bounded by $\mathcal{O}(T^{7/2}/d^2) + \mathcal{O}(T^2/d)$ using Corollary 7.8 in combination with Lemma 10 of [ACQ21], and so in the end we obtain

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{m} \neq \mathbf{1}} |p_{\mathbf{m}, \mathbf{c}}^{\mathcal{S}}(\ell)| \leq \mathcal{O}\left(\frac{T^{7/2}}{d^2}\right) + \mathcal{O}\left(\frac{T^2}{d}\right). \quad (255)$$

Third term for $O(d)$ case

The third term in (232) is $\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{n \neq e, m} |p_{m,n}^{\mathcal{O}}(\ell)|$. Applying the Cauchy-Schwarz inequality we obtain the upper bound

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{n \neq e, m} (dd')^T W_\ell \left| \begin{array}{c} \langle \Phi_\ell | \rightarrow \boxed{\Delta_m} \rightarrow | \Phi_\ell \rangle \\ \langle \Psi_\ell | \rightarrow \boxed{\Delta_n} \leftarrow | \Psi_\ell \rangle \end{array} \right| |\text{Wg}^O(\sigma_n \sigma_m^{-1}, d)|. \quad (256)$$

Here we generalize our notation for the C matrices by writing

$$\begin{aligned}
C(2i-1, 2j) &:= \begin{array}{c} \langle \phi_{v_{i-1}} | \text{---} \text{---} | \phi_{v_{j-1}} \rangle \\ \text{---} \text{---} | \psi_{v_i} \rangle \text{---} \text{---} | \psi_{v_j} \rangle \end{array} \\
C(2i, 2j) &:= \begin{array}{c} \text{---} \text{---} | \phi_{v_{i-1}} \rangle \text{---} \text{---} | \phi_{v_{j-1}} \rangle \\ \text{---} \text{---} | \psi_{v_i} \rangle \text{---} \text{---} | \psi_{v_j} \rangle \end{array}, \quad i < j \\
C(2i-1, 2j-1) &:= \begin{array}{c} \text{---} \text{---} | \phi_{v_{i-1}} \rangle \text{---} \text{---} | \phi_{v_{j-1}} \rangle \\ \text{---} \text{---} | \psi_{v_i} \rangle \text{---} \text{---} | \psi_{v_j} \rangle \end{array}, \quad i < j
\end{aligned}
\tag{257}$$

where as before $C(j, i) := C(i, j)^t$. Then the diagrammatic term in (256) can be written as

$$\left| \text{tr}(C(f_n \circ f_e(1), 1)C(f_m(1), f_n \circ f_e \circ f_m(1))C(f_m \circ f_n \circ f_e \circ f_m(1), f_n \circ f_e \circ f_m \circ f_n \circ f_e \circ f_m(1)) \cdots) \right. \\ \left. \cdots \text{tr}(\cdots) \cdots \text{tr}(\cdots) \right| \quad (258)$$

and so using the same 1-norm bound as before we obtain the upper bound

$$\prod_{i=1}^T \|C(\mathbf{n}(2i), \mathbf{n}(2i-1))\|_1. \quad (259)$$

To simplify (259), we define the unnormalized density operator $\tilde{\rho}_{v_i} \in \text{Mat}_{d \times d}(\mathbb{C})$:

[illegible]

for any $i \in [T]$. Then we have the following Lemma:

Lemma 7.17. *For any $i, j \in [T]$,*

$$\|C(2i-1, 2j)\|_1 \leq \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})} \quad (261)$$

$$\|C(2i, 2j)\|_1 \leq \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t)} \quad (262)$$

$$\|C(2i-1, 2j-1)\|_1 \leq \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t)}. \quad (263)$$

Proof. First consider $\|C(2i-1, 2j)\|_1$. Since $\|A\|_1 = \|A \otimes A^\dagger\|_1^{1/2} \leq \|\text{SWAP} \cdot (A \otimes A^\dagger)\|_1^{1/2}$, we have

$$\left(\begin{array}{cc} \langle \phi_{v_{i-1}} | \rightarrow & \leftarrow | \phi_{v_{j-1}} \rangle \\ \langle \psi_{v_i} | \rightarrow & \leftarrow | \psi_{v_j} \rangle \end{array} \quad \begin{array}{cc} \langle \phi_{v_{j-1}} | \rightarrow & \leftarrow | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_j} | \rightarrow & \leftarrow | \psi_{v_i} \rangle \end{array} \right)^{1/2}.$$

Since the tensor network in the trace is positive semi-definite, we can compute the 1-norm by taking to trace and find

$$\left\| \begin{array}{cc} \langle \phi_{v_{i-1}} | \rightarrow & \leftarrow | \phi_{v_{j-1}} \rangle \\ \langle \psi_{v_i} | \rightarrow & \leftarrow | \psi_{v_j} \rangle \end{array} \right\|_1 \leq \left\| \begin{array}{cc} \langle \phi_{v_{i-1}} | \rightarrow & \leftarrow | \phi_{v_{j-1}} \rangle \\ \langle \psi_{v_i} | \rightarrow & \leftarrow | \psi_{v_j} \rangle \end{array} \quad \begin{array}{cc} \langle \phi_{v_{j-1}} | \rightarrow & \leftarrow | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_j} | \rightarrow & \leftarrow | \psi_{v_i} \rangle \end{array} \right\|_1^{1/2}. \quad (264)$$

We have colored the tensor diagram suggestively so that it is transparent how to rewrite it as $\sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})}$.

Now consider $\|C(2i, 2j)\|_1$. Using the same inequality $\|A\|_1 = \|A \otimes A^\dagger\|_1^{1/2} \leq \|\text{SWAP} \cdot (A \otimes A^\dagger)\|_1^{1/2}$, we find the upper bound

$$\left(\begin{array}{cc} \langle \phi_{v_{i-1}} | \rightarrow & \leftarrow | \phi_{v_{i-1}} \rangle \\ \langle \phi_{v_{j-1}} | \rightarrow & \leftarrow | \phi_{v_{j-1}} \rangle \end{array} \quad \begin{array}{cc} \langle \psi_{v_i} | \rightarrow & \leftarrow | \psi_{v_i} \rangle \\ \langle \psi_{v_j} | \rightarrow & \leftarrow | \psi_{v_j} \rangle \end{array} \right)^{1/2} \quad (265)$$

which is clearly equal to $\sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t)}$. The upper bound on $\|C(2i-1, 2j-1)\|_1$ is given by an identical argument. \square

It will be convenient to define the underline notation

$$\underline{i} := \begin{cases} \frac{i+1}{2} & \text{for } i \text{ odd} \\ \frac{i}{2} & \text{for } i \text{ even} \end{cases} \quad (266)$$

and analogously for \underline{j} . We will say that i, j have same parity if they are equal modulo 2, and have different parity otherwise. Then using the above Lemma, we can upper bound (259) by

$$\prod_{m=1}^{\#O(n)} \left(\prod_{\substack{i \leftrightarrow j \in B_m \\ i, j \text{ opposite parity}}} \sqrt{\text{tr}(\tilde{\rho}_{v_{\underline{i}}} \tilde{\rho}_{v_{\underline{j}}})} \prod_{\substack{k \leftrightarrow \ell \in B_m \\ k, \ell \text{ same parity}}} \sqrt{\text{tr}(\tilde{\rho}_{v_{\underline{k}}} \tilde{\rho}_{v_{\underline{\ell}}})} \right). \quad (267)$$

Now we turn to bounding (267). We begin by developing some notation for cycles B_m . Standard notation for an M_{2T} cycle of “length p ” is $B_m = (a_1 a_2 \cdots a_{2p})$ where for us $\{a_1, a_2, \dots, a_{2p}\} \subseteq [2T]$. Since the number of elements in an M_{2T} -cycle is always even, by convention we define the length as half the number of elements (i.e. p instead of $2p$). We decorate each a_i by an additional subscript

as $a_{m,i}$ to remember that it belongs to the m th cycle B_m . We sometimes write $p = |B_m|$ to remind ourselves that it depends on m . Let us also define

$$\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^{t(i,j)}) := \begin{cases} \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}) & \text{if } i, j \text{ have different parities} \\ \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t) & \text{if } i, j \text{ have the same parity} \end{cases}. \quad (268)$$

With these notations in mind, we can write (267) as a product over

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,1}}} \tilde{\rho}_{v_{a_{m,2}}}^{t(a_{m,1}, a_{m,2})})} \cdots \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,2p-1}}} \tilde{\rho}_{v_{a_{m,2p}}}^{t(a_{m,2p-1}, a_{m,2p})})} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,2p}}} \tilde{\rho}_{v_{a_{m,1}}}^{t(a_{m,2p}, a_{m,1})})} \quad (269)$$

over $m = 1, \dots, \#(\mathbf{n})$; here we drop the O superscript on $\#^O$ since we are only discussing pair permutation here. We will further analyze the above for fixed m in two cases: (i) $p = |B_m|$ is even, and (ii) $p = |B_m|$ is odd.

It will be convenient to prove a Lemma which is slightly more general than what we need for the orthogonal case; the advantage of this generality is that it will immediately apply to the symplectic case. The Lemma is as follows:

Lemma 7.18. *Let $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^{f(i,j)})$ equal either $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})$, $\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t)$, or $\text{tr}(\tilde{\rho}_{v_i} \mathbf{J} \tilde{\rho}_{v_j}^t \mathbf{J}^{-1})$, depending on the value of i, j . Defining*

$$R_{m,-} := \text{tr}(\tilde{\rho}_{v_{a_{m,p}}}) \prod_{\substack{i \text{ odd} \\ 1 \leq i \leq 2p-2}} \text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}^{f(a_{m,i}, a_{m,i+1})}) \quad (270)$$

$$R_{m,+} := \text{tr}(\tilde{\rho}_{v_{a_{m,1}}}) \prod_{j \text{ even}} \text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}}^{f(a_{m,j}, a_{m,j+1})}), \quad (271)$$

we have the inequality

$$\prod_{m=1}^{\#(\mathbf{n})} \prod_{i \leftrightarrow j \in B_m} \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^{f(i,j)})} \leq \frac{1}{2^{\#(\mathbf{n})}} \sum_{i_1, \dots, i_{\#(\mathbf{n})} = \pm} R_{1,i_1} R_{2,i_2} \cdots R_{\#(\mathbf{n}), i_{\#(\mathbf{n})}}. \quad (272)$$

Proof. Similar to the unitary setting, the argument proceeds in two cases.

Case 1: p is even. Every term in Eqn. (269) has the form $\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}^{f(a_{m,i}, a_{m,i+1})})}$ except the last term. The i subscripts of $\tilde{\rho}_{v_{a_{m,i}}}$ will be treated modulo $2p$, and $\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,2p}}} \tilde{\rho}_{v_{a_{m,2p+1}}}^{f(a_{m,2p}, a_{m,2p+1})})}$. With this notation at hand, we organize Eqn. (269) as

$$\left(\prod_{i \text{ odd}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}^{f(a_{m,i}, a_{m,i+1})})} \right) \left(\prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}}^{f(a_{m,j}, a_{m,j+1})})} \right). \quad (273)$$

Since $ab \leq \frac{1}{2}(a^2 + b^2)$, the expression above is upper bounded by

$$\frac{1}{2} \prod_{i \text{ odd}} \text{tr}(\tilde{\rho}_{v_{a_{m,i}}} \tilde{\rho}_{v_{a_{m,i+1}}}^{f(a_{m,i}, a_{m,i+1})}) + \frac{1}{2} \prod_{j \text{ even}} \text{tr}(\tilde{\rho}_{v_{a_{m,j}}} \tilde{\rho}_{v_{a_{m,j+1}}}^{f(a_{m,j}, a_{m,j+1})}). \quad (274)$$

We will call the first term $\frac{1}{2} R_{m,-}$ and the second term $\frac{1}{2} R_{m,+}$.

Case 2: p is odd. In this setting we can arrange the terms in Eqn. (211) as

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,p}} \tilde{\rho}_{v_{a_m,1}}^{f(a_m,p,a_m,1)})} \left(\prod_{\substack{i \text{ odd} \\ 1 \leq i \leq 2p-2}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,i}} \tilde{\rho}_{v_{a_m,i+1}}^{f(a_m,i,a_m,i+1)})} \right) \left(\prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,j}} \tilde{\rho}_{v_{a_m,j+1}}^{f(a_m,j,a_m,j+1)})} \right). \quad (275)$$

Since for A, B Hermitian and positive semi-definite we have $\text{tr}(AB) \leq \|A\|_2 \|B\|_2 \leq \|A\|_1 \|B\|_1 \leq \text{tr}(A) \text{tr}(B)$ it follows that

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,p}} \tilde{\rho}_{v_{a_m,1}}^{f(a_m,p,a_m,1)})} \leq \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,p}}) \text{tr}(\tilde{\rho}_{v_{a_m,1}}^{f(a_m,p,a_m,1)})} = \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,p}}) \text{tr}(\tilde{\rho}_{v_{a_m,1}})}. \quad (276)$$

Using the above inequality, (275) has the upper bound

$$\left(\sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,p}})} \prod_{\substack{i \text{ odd} \\ 1 \leq i \leq 2p-2}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,i}} \tilde{\rho}_{v_{a_m,i+1}}^{f(a_m,i,a_m,i+1)})} \right) \left(\sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,1}})} \prod_{j \text{ even}} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_m,j}} \tilde{\rho}_{v_{a_m,j+1}}^{f(a_m,j,a_m,j+1)})} \right). \quad (277)$$

Further using $ab \leq \frac{1}{2}(a^2 + b^2)$, we find the upper bound

$$\frac{1}{2} \text{tr}(\tilde{\rho}_{v_{a_m,p}}) \prod_{\substack{i \text{ odd} \\ 1 \leq i \leq 2p-2}} \text{tr}(\tilde{\rho}_{v_{a_m,i}} \tilde{\rho}_{v_{a_m,i+1}}^{f(a_m,i,a_m,i+1)}) + \frac{1}{2} \text{tr}(\tilde{\rho}_{v_{a_m,1}}) \prod_{j \text{ even}} \text{tr}(\tilde{\rho}_{v_{a_m,j}} \tilde{\rho}_{v_{a_m,j+1}}^{f(a_m,j,a_m,j+1)}) \quad (278)$$

where similarly the first term is called $\frac{1}{2} R_{m,-}$ and the second term is called $\frac{1}{2} R_{m,+}$.

Since $R_{m,-}$ and $R_{m,+}$ implicitly depend on the leaf ℓ , sometimes we will denote them by $R_{m,-}^\ell$ and $R_{m,+}^\ell$ to be explicit.

Taking Case 1 and Case 2 together, we find that Eqn. (267) has the upper bound

$$\begin{aligned} \prod_{m=1}^{\#(n)} \prod_{i \leftrightarrow j \in B_m} \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^{f(i,j)})} &\leq \frac{1}{2^{\#(n)}} \prod_{m=1}^{\#(n)} (R_{m,-} + R_{m,+}) \\ &= \frac{1}{2^{\#(n)}} \sum_{i_1, \dots, i_{\#(n)} = \pm} R_{1,i_1} R_{2,i_2} \cdots R_{\#(n),i_{\#(n)}}. \end{aligned} \quad (279)$$

This is the desired bound. \square

Observe that each term in the sum in the last line of (279) is a product of terms like $\text{tr}(\tilde{\rho}_{v_i})$ and $\text{tr}(\tilde{\rho}_{v_j} \tilde{\rho}_{v_k}^{f(j,k)})$. As before, the key point is that we arranged the equations so that each term in the sum has $\tilde{\rho}_{v_i}$ for each $i = 1, \dots, T$ appear *exactly once*. This allows us to prove the following lemma, which is akin to Lemma 7.14 above:

Lemma 7.19. *Fix any $i_1, \dots, i_{\#(n)} \in \{+, -\}$. Then*

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell R_{1,i_1}^\ell R_{2,i_2}^\ell \cdots R_{\#(n),i_{\#(n)}}^\ell \leq d^{\frac{T}{2} - \lfloor \frac{L(n)}{2} \rfloor} \quad (280)$$

where $L(\mathbf{n})$ is the length of the longest cycle in \mathbf{n} (where we recall that the length of an M_{2T} cycle is defined as half the number of integers in that cycle).

Proof. To ease notation, let $R_j^\ell \triangleq R_{j,i_j}^\ell$. Recall that $W_\ell = w_{v_1} w_{v_2} \cdots w_{v_T}$ and note that

$$\sum_{v: \text{depth}(v)=i} dd' w_v \tilde{\rho}_v = \langle \phi_{\text{parent}(v)} | \phi_{\text{parent}(v)} \rangle \mathbb{1}_{d \times d} = \mathbb{1}_{d \times d} \quad (281)$$

$$\sum_{v: \text{depth}(v)=i} dd' w_v \tilde{\rho}_v^t = \langle \phi_{\text{parent}(v)} | \phi_{\text{parent}(v)} \rangle \mathbb{1}_{d \times d} = \mathbb{1}_{d \times d} \quad (282)$$

$$\sum_{v: \text{depth}(v)=i} dd' w_v J \tilde{\rho}_v^t J^{-1} = \langle \phi_{\text{parent}(v)} | \phi_{\text{parent}(v)} \rangle \mathbb{1}_{d \times d} = \mathbb{1}_{d \times d}. \quad (283)$$

Taking the traces of the above equations against any $\rho \in \text{Mat}_{d \times d}(\mathbb{C})$, we find

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\rho \tilde{\rho}_v) = \text{tr}(\rho) \quad (284)$$

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\rho \tilde{\rho}_v^t) = \text{tr}(\rho) \quad (285)$$

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\rho J \tilde{\rho}_v^t J^{-1}) = \text{tr}(\rho). \quad (286)$$

In particular, for $\rho = \mathbb{1}$ we have

$$\sum_{v: \text{depth}(v)=i} dd' w_v \text{tr}(\tilde{\rho}_v) = d. \quad (287)$$

With these various identities in mind, we can now turn to bounding (280). As we discussed, above $\prod_j R_j^\ell$ is a product of terms like $\text{tr}(\tilde{\rho}_{v_{\underline{i}}})$ and $\text{tr}(\tilde{\rho}_{v_{\underline{i}}} \tilde{\rho}_{v_{\underline{i}'}}^{f(i,i')})$. It is convenient to define some sets of indices to encode this data. Let $S_1^{(T)} \subseteq [T]$ denote the indices \underline{i} for which $\text{tr}(\tilde{\rho}_{v_{\underline{i}}})$ appears in $\prod_j R_j^\ell$, and also let $S_2^{(T)} \subseteq [T] \times [T]$ denote the set of (unordered) pairs $(\underline{i}, \underline{i}')$ for which $\text{tr}(\tilde{\rho}_{v_{\underline{i}}} \tilde{\rho}_{v_{\underline{i}'}}^{f(i,i')})$ appears, so that for any root-to-leaf path in \mathcal{T} consisting of nodes $v_1, \dots, v_T = \ell$, we have

$$\prod_j R_j^\ell = \prod_{\underline{i} \in S_1^{(T)}} \text{tr}(\tilde{\rho}_{v_{\underline{i}}}) \cdot \prod_{(\underline{i}, \underline{i}') \in S_2^{(T)}} \text{tr}(\tilde{\rho}_{v_{\underline{i}}} \tilde{\rho}_{v_{\underline{i}'}}^{f(i,i')}). \quad (288)$$

Next we construct $S_1^{(t)} \subseteq [t]$ and $S_2^{(t)} \subseteq [t] \times [t]$ for $1 \leq t < T$ by the following inductive procedure. If $t \in S_1^{(t)}$ then we define $S_1^{(t-1)} \triangleq S_1^{(t)} \setminus \{t\}$ and $S_2^{(t-1)} \triangleq S_2^{(t)}$. Otherwise if $(t, t') \in S_2^{(t)}$ for some $t' < t$, then we define $S_1^{(t-1)} \triangleq S_1^{(t-1)} \cup \{t'\}$ and $S_2^{(t-1)} \triangleq S_2^{(t)} \setminus \{(t, t')\}$. We recall two key observations about such sequences, which we also leveraged in Lemma 7.14:

Observation 7.20. For every $i \in S_1^{(T)}$, we have that $i \in S_1^{(i)}$.

Observation 7.21. For every $(i, i') \in S_2^{(T)}$, if $i \leq i'$ then $i \in S_1^{(i)}$ while $i' \notin S_1^{(i')}$.

We have defined this set of sequences in order to extract $\tilde{\rho}_{v_T}$ from the product on the right-hand side of (288) and apply (287) (respectively (284)) if $T \in S_1^{(T)}$ (respectively $(T, t') \in S_2^{(T)}$ for some $t' < T$) to obtain

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell \prod_j R_j^\ell = d^{\mathbb{1}[T \in S_1^{(T)}]} \sum_{u: \text{depth}(u)=T-1} (dd')^{T-1} W_u \prod_{i \in S_1^{(T-1)}} \text{tr}(\tilde{\rho}_{v_i}) \cdot \prod_{(i, i') \in S_2^{(T-1)}} \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_{i'}}^{f(i, i')}), \quad (289)$$

with $W_u = w_{v_1} \cdots w_{v_{T-1}}$ if the path from root to u in \mathcal{T} is given by $v_1, \dots, v_{T-1} = u$. We can proceed inductively by expressing the right-hand side of (289) as

$$d^{\sum_{t=1}^T \mathbb{1}[t \in S_1^{(t)}]}. \quad (290)$$

By virtue of Observations 7.20 and 7.21, $\sum_{t=1}^T \mathbb{1}[t \in S_1^{(t)}] = |S_1^{(T)}| + |S_2^{(T)}|$. Since every M_{2T} -cycle B_m of \mathbf{n} contributes with $|B_m|$ off contributes $|B_m|/2$ pairs to $S_2^{(T)}$, and every M_{2T} -cycle B_m with $|B_m|$ even contributes $\lfloor |B_m|/2 \rfloor$ pairs to $S_2^{(T)}$ and one element to $S_1^{(T)}$, we conclude the formula

$$\begin{aligned} \sum_{\ell \in \text{leaf}(\mathcal{T})} (dd')^T W_\ell \prod_j R_j^\ell &= d^{\sum_{m=1}^{\#(\mathbf{n})} \lceil \frac{|B_m|}{2} \rceil} \\ &= d^{\frac{T}{2} - \sum_{m=1}^{\#(\mathbf{n})} \lfloor \frac{|B_m|}{2} \rfloor} \\ &\leq d^{\frac{T}{2} - \lfloor \frac{L(\mathbf{n})}{2} \rfloor} \end{aligned} \quad (291)$$

as we desired. \square

Putting all our previous bounds together, in particular Eqn.'s (256), (267), (279), and (280), we arrive at

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{n} \neq \mathbf{e}, \mathbf{m}} |p_{\sigma, \tau}^{\mathcal{O}}(\ell)| \leq d^T \sum_{\mathbf{m}} |\text{Wg}^{\mathcal{O}}(\sigma_{\mathbf{m}}^{-1}, d)| \sum_{\mathbf{n} \neq \mathbf{e}} d^{-\lfloor \frac{L^{\mathcal{O}}(\mathbf{n})}{2} \rfloor}. \quad (292)$$

The first sum on the right-hand side is bounded by $d^T \sum_{\mathbf{m}} |\text{Wg}^{\mathcal{O}}(\sigma_{\mathbf{m}}^{-1}, d)| \leq 1 + \mathcal{O}\left(\frac{T^2}{d}\right)$. Considering the second sum on the right-hand side, let $N^{\mathcal{O}}(T, \ell)$ be the number of permutations in S_T where the length of the longest M_{2T} -cycle is 2ℓ . Then the second sum can be written as

$$\sum_{\ell=2}^T N^{\mathcal{O}}(T, \ell) d^{-\lfloor \frac{\ell}{2} \rfloor} \quad (293)$$

where we omit $\ell = 1$ from the sum since it corresponds to the identity pair permutation. Since $N^{\mathcal{O}}(T, \ell) \leq \binom{2T}{2\ell} (2\ell - 1)!! = \frac{(2T)!}{(2T-2\ell)! 2^\ell \ell!} < T^\ell$, (293) is upper bounded by

$$\sum_{\ell=2}^{\infty} T^\ell d^{-\lfloor \frac{\ell}{2} \rfloor} = \frac{(1+T) \frac{T^2}{d}}{1 - \frac{T^2}{d}} = \frac{T^3}{d} + \frac{T^2}{d} + \mathcal{O}\left(\frac{T^5}{d^2}\right). \quad (294)$$

Now if $T \leq o(d^{1/3})$, then this quantity is $o(1)$ for some absolute constant $c > 0$. Altogether, we find

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(\mathcal{T})} \sum_{\mathbf{n} \neq \mathbf{e}, \mathbf{m}} |p_{\mathbf{m}, \mathbf{n}}^{\mathcal{O}}(\ell)| \leq o(1). \quad (295)$$

$\|\text{SWAP} \cdot (A \otimes A^\dagger)\|_1^{1/2}$ to find the upper bound

$$\left(\begin{array}{c} \langle \phi_{v_{j-1}} | \\ \langle \phi_{v_{i-1}} | \\ | \phi_{v_{i-1}} \rangle \\ \langle \psi_{v_j} | \\ \langle \psi_{v_i} | \\ | \psi_{v_i} \rangle \end{array} \right)^{1/2}$$

which evidently equals $\sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^D)}$. □

Leveraging the above Lemma, we can upper bound (299) by

$$\prod_{m=1}^{\#^O(\mathbf{n})} \left(\prod_{\substack{i \leftrightarrow j \in B_m \\ i, j \text{ opposite parity}}} \sqrt{\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j})} \prod_{\substack{k \leftrightarrow \ell \in B_m \\ k, \ell \text{ both odd}}} \sqrt{\text{tr}(\tilde{\rho}_{v_k} \tilde{\rho}_{v_\ell}^t)} \prod_{\substack{p \leftrightarrow q \in B_m \\ p, q \text{ both even}}} \sqrt{\text{tr}(\tilde{\rho}_{v_p} \tilde{\rho}_{v_q}^D)} \right). \quad (303)$$

Defining

$$\text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^{D(i,j)}) := \begin{cases} \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}) & \text{if } i, j \text{ have different parities} \\ \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^t) & \text{if } i, j \text{ are both odd} \\ \text{tr}(\tilde{\rho}_{v_i} \tilde{\rho}_{v_j}^D) & \text{if } i, j \text{ are both even} \end{cases}, \quad (304)$$

we can write (303) as a product over

$$\sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,1}}} \tilde{\rho}_{v_{a_{m,2}}}^{D(a_{m,1}, a_{m,2})})} \cdots \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p-1}}} \tilde{\rho}_{v_{a_{m,p}}}^{D(a_{m,p-1}, a_{m,p})})} \sqrt{\text{tr}(\tilde{\rho}_{v_{a_{m,p}}} \tilde{\rho}_{v_{a_{m,1}}}^{D(a_{m,p}, a_{m,p-1})})} \quad (305)$$

over $m = 1, \dots, \#^{\text{Sp}}(\mathbf{n})$. Leveraging Lemmas 7.18 and 7.19 in the same exact way as in the orthogonal case, we obtain

$$\sum_{\ell \in \text{leaf}(T)} \sum_{\mathbf{n} \neq \mathbf{e}, \mathbf{m}} |p_{\sigma, \tau}^S(\ell)| \leq d^T \sum_{\mathbf{m}} |\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}^{-1}, d/2)| \sum_{\mathbf{n} \neq \mathbf{e}} d^{-\lfloor \frac{L(\mathbf{n})}{2} \rfloor}. \quad (306)$$

The first sum on the right-hand side is bounded by $d^T \sum_{\mathbf{m}} |\text{Wg}^{\text{Sp}}(\sigma_{\mathbf{m}}^{-1}, d/2)| \leq 1 + \mathcal{O}\left(\frac{T^2}{d}\right)$. Considering the second sum on the right-hand side, let $N^{\text{Sp}}(T, \ell)$ be the number of permutations in S_T where the length of the longest M_{2T} -cycle is 2ℓ . Since $N^{\text{Sp}}(T, \ell) = N^O(T, \ell)$, the second sum can be written as

$$\sum_{\ell=2}^T N^{\text{Sp}}(T, \ell) d^{-\lfloor \frac{\ell}{2} \rfloor} \leq \frac{T^3}{d} + \frac{T^2}{d} + \mathcal{O}\left(\frac{T^5}{d^2}\right). \quad (307)$$

Now if $T \leq o(d^{1/3})$, then this quantity is $o(1)$ for some absolute constant $c > 0$. Altogether, we find

$$\frac{1}{2} \sum_{\ell \in \text{leaf}(T)} \sum_{\mathbf{n} \neq \mathbf{e}, \mathbf{m}} |p_{\mathbf{m}, \mathbf{n}}^S(\ell)| \leq o(1). \quad (308)$$

□

7.3.4 Corollaries involving state distinction

Here we remark on some immediate corollaries of Theorems 7.9, 7.10, and 7.11. We begin with a corollary essentially identical to one from [ACQ21]:

Corollary 7.23. *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(2^{n/3}\right), \quad (309)$$

to correctly distinguish between the maximally mixed state $\mathbb{1}/2^n$ on n qubits and a fixed, Haar-random state $|\Psi\rangle\langle\Psi|$ on n qubits with probability at least $2/3$.

Proof. Suppose by contradiction that a learning algorithm could distinguish between $\mathbb{1}/d$ and a fixed, Haar-random $|\Psi\rangle\langle\Psi|$ with probability at least $2/3$ using $T < \mathcal{O}(d^{1/3})$. Then we could use this learning algorithm to solve the unitary distinction problem by taking the channel \mathcal{C} and applying it to $|0\rangle^{\otimes n}$; if $\mathcal{C} = \mathcal{D}$ then we would get the maximally mixed state, and if $\mathcal{C} = \mathcal{U}$ we would get a fixed Haar random state. Moreover we could distinguish the two cases in using $T < \mathcal{O}(d^{1/3})$ by running the alleged state distinction algorithm. But this is impossible since it contradicts Theorem 7.9, so such a state distinction algorithm can not exist. \square

While the above corollary is weaker than Theorem 5.11 for which $T \geq \Omega(d^{1/2})$, it is emblematic of a general strategy for using learning bounds on channel distinction problems to derive corresponding learning bounds on state distinction problems. Further leveraging this strategy, we can prove the following two additional corollaries:

Corollary 7.24. *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(2^{n/3}\right), \quad (310)$$

to correctly distinguish between the maximally mixed state $\mathbb{1}/2^n$ on n qubits and a fixed, real Haar random state $|\Psi\rangle\langle\Psi|$ on n qubits with probability at least $2/3$.

To prove this corollary, we observe that $|\Psi\rangle\langle\Psi| = O(|0\rangle\langle 0|)^{\otimes n} O^t$ for a fixed, Haar-random orthogonal matrix; alternatively, it is also the case that $|\Psi\rangle\langle\Psi| = S(|0\rangle\langle 0|)^{\otimes n} S^D$ for a fixed, Haar-random symplectic matrix. Leveraging the symplectic version (in particular since Theorem 7.10 has a stronger bound than Theorem 7.11), the corollary follows by the same arguments as the proof of Corollary 7.23 in combination with Theorem 7.11. Corollary 7.24 also follow from the results in [ACQ21], although the corollary was not stated there.

7.3.5 Upper bound without quantum memory

The upper bound without quantum memory can be obtained by reducing the problem to a purity testing and utilize Theorem 5.13. This results in the following corollary.

Corollary 7.25. *There is a learning algorithm without quantum memory which takes $T = \mathcal{O}(2^{n/2})$ accesses to the unknown quantum channel \mathcal{C} to distinguish between whether \mathcal{C} is a fixed Haar-random unitary channel or a completely depolarizing channel \mathcal{D} .*

Proof. We perform T repeated experiments given by the following. Input the all-zero state $|0^n\rangle$ to the unknown quantum channel \mathcal{C} to obtain the output state ρ_{out} . When \mathcal{C} is a scrambling unitary channel, ρ_{out} is a fixed pure state. When \mathcal{C} is a completely depolarizing channel, ρ_{out} is

the completely mixed state. We then measure the output state ρ_{out} in the computational basis to obtain the classical data. The collection of classical data given by the computational basis measurements can be used to classify if ρ_{out} is a fixed pure state or the completely mixed state. Theorem 5.13 tells us that $T = O(2^{n/2})$ is sufficient to distinguish between the two cases. Hence, we can distinguish between whether \mathcal{C} is a scrambling unitary channel or a completely depolarizing channel using $T = O(2^{n/2})$. \square

7.3.6 Upper bound with quantum memory

The exponential lower bound for algorithms without quantum memory is in stark contrast to those with quantum memory. The following result from [ACQ21] states that a linear number of channel applications T and quantum gates suffices if we use a learning algorithm with quantum memory.

Theorem 7.26 (Fixed unitary task is easy with an n qubit quantum memory [ACQ21]). *There exists a learning algorithm with n qubits of quantum memory which, with constant probability, can distinguish a completely depolarizing channel \mathcal{D} from a fixed, Haar-random \mathcal{U} (either unitary, orthogonal, or symplectic) using only $T = \mathcal{O}(1)$ applications of the channel. Moreover, the algorithm is gate efficient and has $\mathcal{O}(n)$ gate complexity.*

The protocol is simply a swap test; the basic idea is that for a pure state $|\phi\rangle$ on n qubits we have $\text{tr}(\mathcal{D}[|\phi\rangle\langle\phi|]^2) = \frac{1}{d}$ whereas $\text{tr}(\mathcal{U}[|\phi\rangle\langle\phi|]^2) = 1$. The ability to obtain quantum interference between $\mathcal{D}[|\phi\rangle\langle\phi|]$ and a copy of itself $\mathcal{D}[|\phi\rangle\langle\phi|]$ (or $\mathcal{U}[|\phi\rangle\langle\phi|]$ and a copy of itself $\mathcal{U}[|\phi\rangle\langle\phi|]$) is enabled by the n qubit quantum memory which can store a single copy of the state.

7.4 Symmetry distinction problem

7.4.1 Lower bound

Using Theorem 7.9, 7.10, and 7.11, we can show the hardness of distinguishing between unitary channel, orthogonal matrix channel, and symplectic matrix channel for learning algorithms without quantum memory. This multiple-hypothesis distinguishing task is equivalent to uncovering what symmetry is encoded in a quantum evolution. Orthogonal matrix channel and symplectic matrix channel are quantum evolutions with different type of time-reversal symmetry, while unitary channel is a general evolution without additional symmetry.

Theorem 7.27. *Any learning algorithm without quantum memory requires*

$$T \geq \Omega\left(2^{2n/7}\right), \quad (311)$$

to correctly distinguish between a fixed, Haar-random unitary channel \mathcal{C}^U , orthogonal matrix channel \mathcal{C}^O , or symplectic matrix channel \mathcal{C}^S on n qubits with probability at least $2/3$.

Proof. Given a tree representation \mathcal{T} of the learning algorithm without quantum memory. The probability that the algorithm correctly identifies the class of channels is equal to

$$\frac{1}{3} \sum_{\mathcal{C} \in \{\mathcal{C}^U, \mathcal{C}^O, \mathcal{C}^S\}} \sum_{\ell \in \text{leaf}(\mathcal{T})} p^{\mathcal{C}}(\ell) \mathbb{I}[\Upsilon(\ell) = \mathcal{C}], \quad (312)$$

where $\Upsilon(\ell)$ is an element in the set $\{\mathcal{C}^U, \mathcal{C}^O, \mathcal{C}^S\}$ equal to the output when the algorithm arrives at the leaf ℓ . It is not hard to see that the success probability is upper bounded by

$$\frac{1}{3} \sum_{\ell \in \text{leaf}(\mathcal{T})} \max(p^{\mathcal{C}^U}(\ell), p^{\mathcal{C}^O}(\ell), p^{\mathcal{C}^S}(\ell)) \quad (313)$$

$$\leq \frac{1}{3} \sum_{\ell \in \text{leaf}(\mathcal{T})} p^{\mathcal{D}}(\ell) + \max_{\mathcal{C} \in \{\mathcal{C}^U, \mathcal{C}^O, \mathcal{C}^S\}} (|p^{\mathcal{C}}(\ell) - p^{\mathcal{D}}(\ell)|) \quad (314)$$

$$\leq \frac{1}{3} + \frac{1}{3} \sum_{\mathcal{C} \in \{\mathcal{C}^U, \mathcal{C}^O, \mathcal{C}^S\}} \left(\sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{C}}(\ell) - p^{\mathcal{D}}(\ell)| \right). \quad (315)$$

From the proof of Theorem 7.9, 7.10, and 7.11, we have when $T = o(2^{2n/7})$, we have

$$\sum_{\ell \in \text{leaf}(\mathcal{T})} |p^{\mathcal{C}}(\ell) - p^{\mathcal{D}}(\ell)| = o(1), \quad \forall \mathcal{C} \in \{\mathcal{C}^U, \mathcal{C}^O, \mathcal{C}^S\}. \quad (316)$$

Therefore, when $T = o(2^{2n/7})$, the success probability will be upper bounded by $1/3 + o(1)$. Hence to achieve a success probability of at least $2/3$, we must have $T = \Omega(2^{2n/7})$. \square

7.4.2 Upper bound

We present an upper bound for algorithms without quantum memory. There is still a gap between the lower and upper bounds, which we leave as an open question.

Theorem 7.28. *There is an algorithm without quantum memory that uses*

$$T = \mathcal{O}(2^n), \quad (317)$$

to correctly distinguish between a fixed, Haar-random unitary channel \mathcal{C}^U , orthogonal matrix channel \mathcal{C}^O , or symplectic matrix channel \mathcal{C}^S on n qubits with probability at least $2/3$.

We will perform three sets of experiments. Each set of experiments is a quantum state tomography based on random Clifford measurements [KRT17b, GKKT20b] on the output state $\mathcal{C}(|\psi_i\rangle\langle\psi_i|)$ for an input pure state $|\psi_i\rangle$, for $i = 1, 2, 3$. We will take $|\psi_1\rangle = |0^n\rangle$, $|\psi_2\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \otimes |0^{n-1}\rangle$, and $|\psi_3\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle) \otimes |0^{n-1}\rangle$. Note that for each i , $\mathcal{C}(|\psi_i\rangle\langle\psi_i|)$ is a pure state, which we will denote by $\phi_i \in \mathbb{C}^{2^n}$.

We will use the following guarantee:

Lemma 7.29 (State tomography, see [KRT17b, GKKT20b]). *There is an algorithm which for any $\varepsilon > 0$, given copies of an unknown pure state $|\phi\rangle$ with density matrix $\rho \in \mathbb{H}^{2^n \times 2^n}$, makes $\mathcal{O}(2^n/\varepsilon^2)$ random Clifford measurements and outputs the density matrix $\hat{\rho}$ of some pure state for which $\|\rho - \hat{\rho}\|_{\text{tr}} \leq \varepsilon$ with probability at least $14/15$.*

Corollary 7.30. *Suppose $\hat{\rho}_i = |\hat{\phi}_i\rangle\langle\hat{\phi}_i|$ is the output of applying the algorithm in Lemma 7.29 to $\rho = |\phi_i\rangle\langle\phi_i|$. Then provided the algorithm succeeds, we have that for any matrix $M \in \mathbb{C}^{2^n \times 2^n}$ and any $i, j \in \{1, 2, 3\}$*

$$\left| \left| \hat{\phi}_i^t M \hat{\phi}_j \right| - \left| \phi_i^t M \phi_j \right| \right| \leq 2\varepsilon \|M\|_{\infty}. \quad (318)$$

Proof. Because $\|\rho_i - \hat{\rho}_i\|_F \leq \|\rho_i - \hat{\rho}_i\|_{\text{tr}} \leq \varepsilon$, we have that $\|\phi_i - \zeta \cdot \hat{\phi}_i\| \leq \varepsilon$ for some choice of phase $\zeta_i \in \mathbb{C}$. As $|\zeta_i \zeta_j \cdot \hat{\phi}_i^t M \hat{\phi}_j| = |\hat{\phi}_i^t M \hat{\phi}_j|$, we may assume without loss of generality that $\zeta_i, \zeta_j = 1$. Now note that

$$|\hat{\phi}_i^t M \hat{\phi}_j| \leq |\phi_i^t M \hat{\phi}_j| + \varepsilon \|M\|_\infty \leq |\phi_i^t M \phi_j| + 2\varepsilon \|M\|_\infty \quad (319)$$

by triangle inequality, from which the claim follows. \square

Henceforth, let $\hat{\phi}_i$ denote the pure state obtained by applying state tomography to copies of ϕ_i with $\varepsilon = 1/5$. We collect the following basic facts about $\hat{\phi}_i$ which will allow us to distinguish among the three types of channels.

Lemma 7.31. *Condition on the outcome of Lemma 7.29. If \mathcal{C} is a Haar-random symplectic matrix channel, then $|\hat{\phi}_2^t J \hat{\phi}_3| > 1/2$. If \mathcal{C} is a Haar-random unitary or orthogonal channel, then $|\hat{\phi}_2^t J \hat{\phi}_3| < 1/2$ with probability at least 14/15 over the randomness of the channel.*

Proof. By definition of the symplectic matrix channel, $\phi_2^t J \phi_3 = \psi_2^t S^t J S \psi_3 = \psi_2^t J \psi_3 = -1$. The first part of the lemma follows by Corollary 7.30 and the fact that $\|J\|_\infty = 1$. For the second part, note that the joint distribution on (ϕ_2, ϕ_3) is invariant under the transformation $(\phi_2, \phi_3) \mapsto (\phi_2, -\phi_3)$ when \mathcal{C} is a Haar-random unitary (resp. orthogonal) transformation, because conditioned on ϕ_2 , ϕ_3 is a Haar-random in the subspace in \mathbb{C}^{2n} (resp. \mathbb{R}^{2n}) orthogonal to ϕ_2 . So in either case,

$$\mathbb{E}[\phi_2^t J \phi_3] = \frac{1}{2} \mathbb{E}[\phi_2^t J \phi_3] + \mathbb{E}[\phi_2^t J (-\phi_3)] = 0. \quad (320)$$

Note that the function $F : O \mapsto \psi_2^t O^t J O \psi_3$ is 2-Lipschitz:

$$|F(O_1) - F(O_2)| \leq |\psi_2^t (O_1 - O_2)^t J O_1 \psi_3| + |\psi_2^t O_2^t J (O_1 - O_2) \psi_3| \leq 2 \|O_1 - O_2\|_F. \quad (321)$$

So by Lemma 4.3, with probability at least 9/10 over the randomness of the channel, we have that $|\phi_2^t J \phi_3| \leq \mathcal{O}(1/2^{n/2})$. The second part of the lemma then follows from Corollary 7.30. \square

Lemma 7.32. *Condition on the outcome of Lemma 7.29. If \mathcal{C} is a Haar-random orthogonal matrix channel, then $|\hat{\phi}_1^t \hat{\phi}_1| = 1$. If \mathcal{C} is a Haar-random unitary matrix channel, then $|\hat{\phi}_1^t \hat{\phi}_1| < 1/2$ with probability at least 14/15 over the randomness of the channel.*

Proof. Because ϕ_1 is a unit vector with only real entries, $\hat{\phi}_1^t \hat{\phi}_1 = 1$, so the first part of the claim follows by Corollary 7.30. For the second part, if the channel is Haar-random unitary, then ϕ_1 is a Haar-random complex unit vector, so $\mathbb{E}[\hat{\phi}_1^t \hat{\phi}_1] = 0$. The function $F : U \mapsto \psi_1^t U^t U \psi_1$ is 2-Lipschitz by a calculation completely analogous to (321). So by Lemma 4.3, with probability at least 9/10 over the randomness of the channel, we have that $|\hat{\phi}_1^t \hat{\phi}_1| \leq \mathcal{O}(1/2^{n/2})$. The second part of the lemma then follows from Corollary 7.30. \square

We are now ready to prove Theorem 7.28.

Proof of Theorem 7.28. The algorithm will be to apply the state tomography algorithm of Lemma 7.29 to the outputs of $\{|\psi_i\rangle\}$ under the channel, yielding pure states $\{|\hat{\phi}_i\rangle\}$. By a union bound, with probability 2/3 we have that the state tomography algorithm succeeds for all $i = 1, 2, 3$, and Lemmas 7.31 and 7.32 hold. We form the quantity $|\hat{\phi}_2^t J \hat{\phi}_3|$ and check whether it exceeds 1/2. If so, we conclude by Lemma 7.31 that \mathcal{C} is symplectic. Otherwise, we form the quantity $|\hat{\phi}_1^t \hat{\phi}_1|$ and check whether it exceeds 1/2. If so, we conclude by Lemma 7.32 that \mathcal{C} is orthogonal, otherwise it is unitary. \square

References

- [Aar18] Scott Aaronson. Shadow tomography of quantum states. In *STOC*, pages 325–338, 2018.
- [ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The non-stochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- [ACQ21] Dorit Aharonov, Jordan Cotler, and Xiao-Liang Qi. Quantum algorithmic measurement. *arXiv:2101.04634*, 2021.
- [AGZ09] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.
- [ANSV08] Koenraad MR Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, 2008.
- [AR19] Scott Aaronson and Guy N Rothblum. Gentle measurement of quantum states and differential privacy. In *STOC*, pages 322–333, 2019.
- [BC09] Stephen M Barnett and Sarah Croke. Quantum state discrimination. *Advances in Optics and Photonics*, 1(2):238–278, 2009.
- [BC17] Jacob C Bridgeman and Christopher T Chubb. Hand-waving and interpretive dance: an introductory course on tensor networks. *Journal of physics A: Mathematical and theoretical*, 50(22):223001, 2017.
- [BCG13] Konrad Banaszek, Marcus Cramer, and David Gross. Focus on quantum tomography. *New Journal of Physics*, 15(12):125020, 2013.
- [BCHJ⁺21] Fernando G.S.L. Brandão, Wissam Chemissany, Nicholas Hunter-Jones, Richard Kueng, and John Preskill. Models of quantum complexity growth. *PRX Quantum*, 2(3):030316, 2021.
- [BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 692–703. IEEE, 2020.
- [BG18] Sébastien Bubeck and Shirshendu Ganguly. Entropic clt and phase transition in high-dimensional wishart matrices. *International Mathematics Research Notices*, 2018(2):588–606, 2018.
- [BK10] Robin Blume-Kohout. Optimal, reliable estimation of quantum states. *New Journal of Physics*, 12(4):043034, 2010.
- [BKL⁺19] Fernando G.S.L. Brandão, Amir Kalev, Tongyang Li, Cedric Yen-Yu Lin, Krysta M Svore, and Xiaodi Wu. Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. In *ICALP*, 2019.

- [BO20] Costin Bădescu and Ryan O’Donnell. Improved quantum data analysis. *arXiv:2011.10908*, 2020.
- [BOW19] Costin Bădescu, Ryan O’Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 503–514, 2019.
- [CDGR18] Clément L Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. *Theory of Computing Systems*, 62(1):4–62, 2018.
- [CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.
- [Che00] Anthony Chefles. Quantum state discrimination. *Contemporary Physics*, 41(6):401–424, 2000.
- [Chr06] Matthias Christandl. *The structure of bipartite quantum states: insights from group theory and cryptography*. PhD thesis, University of Cambridge, 2006.
- [CLO21] Sitan Chen, Jerry Li, and Ryan O’Donnell. Toward instance-optimal state certification with incoherent measurements. *arXiv:2102.13098*, 2021.
- [CM17] Benoît Collins and Sho Matsumoto. Weingarten calculus via orthogonality relations: new applications. *arXiv:1701.04493*, 2017.
- [CvSW⁺20] Ophelia Crawford, Barnaby van Straaten, D. Wang, T. Parks, E. Campbell, and S. Brierley. Efficient quantum measurement of pauli operators in the presence of finite sampling error. *arXiv:1908.06942*, 2020.
- [CW20] Jordan Cotler and Frank Wilczek. Quantum overlapping tomography. *Phys. Rev. Lett.*, 124(10):100401, 2020.
- [CYZF20] Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T Flammia. Robust shadow estimation. *arXiv:2011.09636*, 2020.
- [CZSJ21] Senrui Chen, Sisi Zhou, Alireza Seif, and Liang Jiang. Quantum advantages for pauli channel estimation. *arXiv:2108.08488*, 2021.
- [DKN14] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 1841–1854. SIAM, 2014.
- [Dys62] Freeman J Dyson. The threefold way. algebraic structure of symmetry groups and ensembles in quantum mechanics. *Journal of Mathematical Physics*, 3(6):1199–1215, 1962.
- [EAO⁺02] Artur K Ekert, Carolina Moura Alves, Daniel KL Oi, Michał Horodecki, Paweł Horodecki, and Leong Chuan Kwek. Direct estimations of linear and nonlinear functionals of a quantum state. *Physical Review Letters*, 88(21):217901, 2002.
- [EHF19] Tim J Evans, Robin Harper, and Steven T Flammia. Scalable bayesian hamiltonian learning. *arXiv:1912.07636*, 2019.

- [FO21] Steven T Flammia and Ryan O’Donnell. Pauli error estimation via population recovery. *arXiv:2105.02885*, 2021.
- [FW20] Steven T Flammia and Joel J Wallman. Efficient estimation of pauli channels. *ACM Transactions on Quantum Computing*, 1(1):1–32, 2020.
- [GAD⁺19] Pranav Gokhale, Olivia Angiuli, Yongshan Ding, Kaiwen Gui, Teague Tomesh, Martin Suchara, Margaret Martonosi, and Frederic T Chong. Minimizing state preparations in variational quantum eigensolver by partitioning into commuting families. *arXiv:1907.13623*, 2019.
- [GKKT20a] M Guță, Jonas Kahn, Richard Kueng, and Joel A Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, 2020.
- [GKKT20b] M Guță, Jonas Kahn, Richard Kueng, and Joel A Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, 2020.
- [GLF⁺10] David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical Review Letters*, 105(15):150401, 2010.
- [Gu13] Yinzheng Gu. *Moments of random matrices and weingarten functions*. PhD thesis, Queen’s University, 2013.
- [Har13] Aram W. Harrow. The church of the symmetric subspace. *arXiv:1308.6595*, 2013.
- [HFW20] Robin Harper, Steven T Flammia, and Joel J Wallman. Efficient learning of quantum noise. *Nature Physics*, 16(12):1184–1188, 2020.
- [HHJ⁺17] Jeongwan Haah, Aram W Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Inf. Theory*, 63(9):5628–5641, 2017.
- [HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.
- [HKP21a] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Efficient estimation of pauli observables by derandomization. *Phys. Rev. Lett.*, 127 3:030503, 2021.
- [HKP21b] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.*, 126:190505, May 2021.
- [HKT⁺21] Hsin-Yuan Huang, Richard Kueng, Giacomo Torlai, Victor V Albert, and John Preskill. Provably efficient machine learning for quantum many-body problems. *arXiv:2106.12627*, 2021.
- [HMR⁺19] William J Huggins, Jarrod McClean, Nicholas Rubin, Zhang Jiang, Nathan Wiebe, K Birgitta Whaley, and Ryan Babbush. Efficient and noise resilient measurements for quantum chemistry on near-term quantum computers. *arXiv:1907.13117*, 2019.

- [Hra97] Zdenek Hradil. Quantum-state estimation. *Physical Review A*, 55(3):R1561, 1997.
- [IS12] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169. Springer Science & Business Media, 2012.
- [KMT⁺17] Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, 549(7671):242–246, 2017.
- [KRT17a] Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080, 2017.
- [KRT17b] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.
- [Lan12] Joseph M Landsberg. Tensors: geometry and applications. *Representation theory*, 381(402):3, 2012.
- [Mat13] Sho Matsumoto. Weingarten calculus for matrix ensembles associated with compact symmetric spaces. *arXiv:1301.5401*, 2013.
- [MdW13] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *arXiv:1310.2035*, 2013.
- [MM17] Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566. PMLR, 2017.
- [MM18] Dana Moshkovitz and Michal Moshkovitz. Entropy samplers and strong generic lower bounds for space bounded learning. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- [MRBAG16] Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New J. Phys.*, 18(2):023023, 2016.
- [MRL08] Masoud Mohseni, Ali T Rezakhani, and Daniel A Lidar. Quantum-process tomography: Resource analysis of different strategies. *Physical Review A*, 77(3):032322, 2008.
- [Orú19] Román Orús. Tensor networks for complex quantum systems. *Nature Reviews Physics*, 1(9):538–550, 2019.
- [OW15] Ryan O’Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 529–538, 2015.
- [OW16] Ryan O’Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 899–912, 2016.
- [OW17] Ryan O’Donnell and John Wright. Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pages 962–974, 2017.

- [Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.
- [PK19] Marco Painsi and Amir Kalev. An approximate description of quantum states. *arXiv:1910.10543*, 2019.
- [Raz17] Ran Raz. A time-space lower bound for a large class of learning problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 732–742. IEEE, 2017.
- [Raz18] Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. *Journal of the ACM (JACM)*, 66(1):1–18, 2018.
- [RYCS21] Zane M Rossi, Jeffery Yu, Isaac L Chuang, and Sho Sugiura. Quantum advantage for noisy channel discrimination. *arXiv:2105.08707*, 2021.
- [SSV19] Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 890–901, 2019.
- [SVW16] Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516. PMLR, 2016.
- [TBG17] Kristan Temme, Sergey Bravyi, and Jay M Gambetta. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, 119(18):180509, 2017.
- [Ver18] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.
- [WGFC14] Nathan Wiebe, Christopher Granade, Christopher Ferrie, and David G Cory. Hamiltonian learning and certification using quantum resources. *Physical Review Letters*, 112(19):190501, 2014.
- [WPS⁺17] Jianwei Wang, Stefano Paesani, Raffaele Santagati, Sebastian Knauer, Antonio A Gentile, Nathan Wiebe, Maurangelo Petruzzella, Jeremy L O’Brien, John G Rarity, Anthony Laing, et al. Experimental quantum hamiltonian learning. *Nature Physics*, 13(6):551–555, 2017.
- [Wri16] John Wright. *How to learn a quantum state*. PhD thesis, Carnegie Mellon University, 2016.
- [Yu97] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.
- [YVI19] Tzu-Ching Yen, Vladyslav Verteletsky, and Artur F Izmaylov. Measuring all compatible operators in one series of a single-qubit measurements using unitary transformations. *arXiv:1907.09386*, 2019.