MATH595: Quantum Learning Theory

Jacob Beckey

Spring 2026

Jacob Beckey
MATH595: Quantum Learning Theory
Spring 2026

University of Illinois, Urbana-Champaign
Department of Mathematics
1409 W Green St
Urbana, IL 61801

Acknowledgements

These notes are an expanded version of what was covered in a one-semester graduate special topics course at the University of Illinois, Urbana-Champaign. I am indebted to the Department of the Mathematics at the University of Illinois for allowing me to create this course and to the students for their feedback that improved these notes. I am also grateful to the various researchers who have taught some of the first courses on this topic:

- John Wright's course offered during the Fall 2024 semester at Berkeley
- Sitan Chen and Jordan Cotler's course offered during Fall 2025 at Harvard
- Robert Huang's course offered during Fall 2025 at Caltech

I have learned a great deal from their excellent lecture notes and, their research in this area generally. It is a very exciting time for this burgeoning field and I hope these notes will be useful for students hoping to learn about learning in the quantum realm!

Contents

1 Introduction

> *What we observe is not nature itself, but nature exposed to our method of questioning.*
>
> — **Werner Heisenberg**

I do not think I have anything terribly unique to add to a crash course on quantum mechanics and quantum information theory, so I refer anyone that needs a refresher to Ref. [NC00]. Once you have learned all of quantum infromation and computing, please come back and learn some quantum laerning theory!

1.1 What is Quantum Learning Theory?

To be written...

Quantum State Discrimination

2

> *The idea of distinguishing probability distributions is slippery business.*
>
> — **Chris Fuchs**

We will begin this course with a topic that is fundamental not only to quantum learning theory, but to quantum mechanics itself: distinguishing quantum states. In addition to being a philosophically interesting topic, state distinguishability also allows us to introduce many useful concepts we will use throughout the course: quantum measurements, distance measures on the space of quantum states, distances between classical probability distributions, and concentration inequalities from classical statistics. Whats more, a standard technique for proving sample complexity lower bounds will involve reducing quantum state discrimination to a problem of interest. All this is to say: pay attention! This stuff is important. Lets begin with the simplest case of discriminating two pure quantum states.[1]

2.1 Pure State Discrimination

Our starting point is simple to state, easy to visualize, and conceptually rich.

Problem 2.1.1 (Pure State Discrimination)**.** Given a pure quantum state $|\psi\rangle \in \mathbb{C}^d$, which is promised to either be $|\psi_1\rangle$ or $|\psi_2\rangle$, determine which is the case.

We assume that the learner has the full classical descriptions of $|\psi_1\rangle$ and $|\psi_2\rangle$ and can use this information to perform any allowed quantum measurement on the unknown state $|\psi\rangle$, regardless of how experimentally feasible this measurement is. Thus, we refer to this as an *information-theoretic* problem, because we are not concerned with the computational or experimental efficiency of whatever strategy we cook up.

Note, it has been understood since the first mathematical formalizations of quantum mechanics that quantum states with non-zero overlap cannot be perfectly

[1] These initial lectures follow along the lines of John Wright's notes at Berkeley [Wri24].

distinguished [Dir58; Neu55]. However, it was not properly formalized in decision-theoretic language until nearly many decades later (see below).

2.1.1 Review: Quantum Measurements

Recall that the most general measurements allowed by quantum mechanics are given by *positive operator-valued measures* (POVMs), defined as follows.

Definition 2.1.2 (Positive operator-valued measure (POVM))**.** A *positive operator-valued measure* (POVM) is a collection of operators $\{E_i\}_i$ satisfying

(i) Positivity $E_i \geq 0$,

(ii) Completeness: $\sum E_i = \mathbb{I}_{\mathcal{H}}$. If we measure a state $\rho \in \mathcal{H}$, we obtain the outcome "i" with probability $p_i = \mathrm{tr}\left[\rho E_i\right]$.

These two properties ensure that the collection of real numbers $\{p_i\}$ forms a probability distribution. We know that $\rho \geq 0$, thus by the spectral theorem, we can always write $\rho = \sum_{j=1} \lambda_j |\psi_j\rangle\langle\psi_j|$, which allows us to write

$$p_i = \mathrm{tr}\left[\rho E_i\right] = \sum_{j=1} \lambda_j \mathrm{tr}\left[|\psi_j\rangle\langle\psi_j| E_i\right] = \sum_{j=1} \lambda_j \langle\psi_j| E_i |\psi_j\rangle . \tag{2.1}$$

Because $\rho \geq 0$, we know $\lambda_j \geq 0$. Thus, for p_i to be non-negative, we require $\langle\psi_j| E_i |\psi_j\rangle \geq 0$ for all possible $|\psi_j\rangle \in \mathbb{C}^d$. This is precisely the definition of being a positive semi-definite operator, thus we see why the positivity assumption is needed. A valid probability distribution must also be normalized

$$p_i = \sum_i \mathrm{tr}\left[\rho E_i\right] = \mathrm{tr}\left[\rho \left(\sum_i E_i\right)\right] = \mathrm{tr}\left[\rho\right] = 1, \tag{2.2}$$

because density operators have unit trace. A very important subset of POVMs are so-called *projection-valued measures* or PVMs that project our density matrix onto a particular subspace.

Definition 2.1.3 (Projection-valued Measure (PVM))**.** A *projective measurement* is a POVM $\{\Pi_i\}_{i=1}^m$ such that $\Pi_i \Pi_j = \delta_{ij} \Pi_i$. Measuring $\rho \in \mathcal{D}(\mathbb{C}^d)$ will yield outcome "i" with probability

$$p_i = \mathrm{tr}\left[\rho \Pi_i\right] . \tag{2.3}$$

In general, these orthogonal projectors can be expressed as

$$\Pi_i = \sum_{j=1}^{r_i} |v_{i,j}\rangle\langle v_{i,j}|, \tag{2.4}$$

where the set $\{|v_{i,j}\rangle\}_{i\in[m],j\in[r_i]}$ forms an orthonormal basis and r_i is the rank of Π_i. The simplest, but most restrictive class of POVMs are obtained when we restrict all Π_i's forming a PVM to be rank-1. We then call this a *basis measurement*.

Definition 2.1.4 (Basis measurement)**.** Let $\{|v_i\rangle\}_{i=1}^d$ be an orthonormal basis of $\mathbb{C}^d$. A *basis measurement* is a PVM $\{\Pi_i\}_{i=1}^d$, where $\Pi_i = |v_i\rangle\langle v_i|$. Given $\rho \in \mathcal{D}(\mathbb{C}^d)$, a basis measurement yields outcome "i" with probability

$$p_i = \mathrm{tr}\left[\rho |v_i\rangle\langle v_i|\right] = \langle v_i|\,\rho\,|v_i\rangle\,. \tag{2.5}$$

In particular, a *standard basis measurement* refers to a basis measurement with respect to the standard basis of $\mathbb{C}^d$, i.e. $\{|i\rangle\}_{i=1}^d$. One of the main themes of this course will be understanding how the allowed class of measurements affects the sample, memory, or time complexity of various learning and testing protocols. See Exercise 2.1.1 for details on simulating all of these measurements using only PVMs.

Quick Quiz 2.1.5. Of these measurement classes, which are repeatable? That is if I measure and obtain outcome "i", which are guaranteed to give outcome "i" if measured again immediately? Which class guarantees a pure post-measurement state?

2.1.2 Figure of Merit for State Discrimination

With these definitions in place, we may now simply state that our allowable strategies are simply a POVM followed by a guess $g \in \{1, 2\}$.

Quick Quiz 2.1.6. Do we need to consider POVMs containing an arbitrary number of elements for this discrimination task?

The task, as we have set it up, requires a definite answer, thus regardless of how many POVM elements we use, we have to define a rule that maps all outcomes to either $g = 1$ or $g = 2$. This process is called *coarse-graining*. It is a useful simplification that will make the error analysis more straightforward.

Without loss of generality, then, a discrimination strategy is described by a two-outcome POVM $E = \{E_1, E_2\}$. If we observe outcome 1, we guess the state $|\psi_1\rangle$ and

if we observe outcome 2, we guess the state $|\psi_2\rangle$. If the actual underlying state is $|\psi\rangle = |\psi_1\rangle$, then the probability of error is given as

$$\Pr[\text{Guess 2}|\text{State 1}] = \operatorname{tr}\left[E_2|\psi_1\rangle\langle\psi_1|\right]. \tag{2.6}$$

By the same logic, if the underlying state is $|\psi\rangle = |\psi_2\rangle$, then an error occurs with probability

$$\Pr[\text{Guess 1}|\text{State 2}] = \operatorname{tr}\left[E_1|\psi_2\rangle\langle\psi_2|\right]. \tag{2.7}$$

When we have no prior information on what state we will be given, it is natural to minimize the worst-case error defined as follows.

Definition 2.1.7 (Worst-case error)**.** For a given measurement strategy defined by a POVM $\{E_1, E_2\}$, the **worst-case error** is the larger of the two conditional error probabilities:

$$P_{\text{worst}} = \max\left\{\Pr[\text{Guess 1} \mid \text{State 2}], \Pr[\text{Guess 2} \mid \text{State 1}]\right\} \tag{2.8}$$

As stated above, our goal is to find the strategy that *minimizes* this *maximum* error. The resulting optimal value is referred to as the **minimax error**:

$$P_{\text{minimax}} = \min_{\{E_1,E_2\}} \max\left\{\operatorname{tr}\left[E_1|\psi_2\rangle\langle\psi_2|\right], \operatorname{tr}\left[E_2|\psi_1\rangle\langle\psi_1|\right]\right\}. \tag{2.9}$$

To understand this set-up operationally, suppose there is an all-knowing referee, Eve, that will prepare the unknown state $|\psi\rangle$ for us. In this course, keeping with the tradition in quantum information theory, we will let Alice and Bob be the agents trying to discriminate, learn, test, communicate, etc. In this case, we only need to introduce Alice as the agent attempting to discriminate these states.

Suppose Alice decides she is just always going to answer $|\psi_1\rangle$. Then Eve, adversarially, will prepare $|\psi\rangle = |\psi_2\rangle$ to ensure Alice is wrong as often as possible. It will be helpful to use this framing to think through our various strategies.

2.1.3 Trivial Strategies and Limiting Cases

Okay, so the stage is set: Alice needs to decide on a strategy for discriminating $|\psi_1\rangle$ and $|\psi_2\rangle$ given only one copy of the unknown state $|\psi\rangle$. Moreover, Eve can adversarially prepare $|\psi\rangle$ after seeing Alice's strategy.

Quick Quiz 2.1.8. Can you guess the functional form of the success probability for pure state discrimination?

At first glance, this may seem intractable, but it turns out to be rather straightforward once we consider some trivial strategies and limiting cases.

Trivial Deterministic Strategy: Always pick the same state. First, consider perhaps the most trivial strategy: always guess $|\psi_1\rangle$ (or, $|\psi_2\rangle$... it doesn't matter). In this case, Eve can just prepare the opposite state and ensure Alice is incorrect with probability 1. This is as bad as it gets!

Trivial Probabilistic Strategy: Flip a coin! Now, suppose Alice has a fair coin at her disposal. She isn't sure how to outsmart Eve, so she decides she is just going to flip this coin and guess the state accordingly. How does this strategy fare in the worst-case error setting? Well, regardless of what Eve does, Alice will be correct half the time (i.e. the worst-case error will be 1/2). Although this is not terribly clever, it is useful in that it gives us a *non-trivial lower bound on the error probability*. We can always achieve a worst-case error probability of 1/2. This provides our benchmark that any non-trivial strategy must improve upon.

Limiting case: identical states[2]. In fact, this strategy is optimal in one case: when $|\psi_1\rangle = |\psi_2\rangle$. When our two states are actually the same state, we might as well just flip a coin and guess randomly. There is no measurement in the universe that can tell us which index Eve chose.

Limiting case: orthogonal states. The other limiting case is when $\langle\psi_1|\psi_2\rangle = 0$. When the two states are guaranteed to be orthogonal, we can distinguish them with unit probability by simply measuring in a basis containing the states. Thus, our success probability should interpolate smoothly between these two extremes.

Given these observations, we might make an educated guess that the optimal success probability is

$$p_{\text{succ}}(\theta) = \frac{1}{2} + \frac{1}{2}\sin\theta, \tag{2.10}$$

where θ is taken to be the (Hilbert space) angle between $|\psi_1\rangle$ and $|\psi_2\rangle$. This seems to work with our limiting cases, so we should have some confidence in this conjectured form! Now that we have thought like physicists, its time to think like mathematicians.

[2] In this scenario, imagine Eve has two identical state preparation machines that are labeled, so she knows which machine produced the state.

2.1.4 Optimal Strategy for Pure State Discrimination

As stated, Problem 2.1.1 involves distinguishing two vectors in an arbitrarily large, but finite, dimensional vector space. This seems daunting until we realize it suffices to consider the subspace spanned by $|\psi_1\rangle, |\psi_2\rangle$.

Dimensional reduction: $\mathbb{C}^d \to \mathbb{C}^2$. Because we are promised that the state is either $|\psi_1\rangle$ or $|\psi_2\rangle$, we know that the unknown state $\rho = |\psi\rangle\langle\psi|$ must lie in the two-dimensional subspace $\mathcal{S} = \text{span}\{|\psi_1\rangle, |\psi_2\rangle\}$. Let, $\Pi_{\mathcal{S}}$ denote the orthogonal projector onto this subspace. Now, suppose we have a strategy that utilizes a POVM acting non-trivially on all of $\mathbb{C}^d$. The measurement statistics will be given as

$$\begin{aligned} \text{tr}\left[E\rho\right] &= \text{tr}\left[E \cdot \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}}\right], \quad \rho = \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}} && (2.11)\\ &= \text{tr}\left[\Pi_{\mathcal{S}}E\Pi_{\mathcal{S}} \cdot \rho\right], \quad \text{cyclicity of trace} && (2.12)\\ &= \text{tr}\left[E'\rho\right], && (2.13) \end{aligned}$$

where $E' := \Pi_{\mathcal{S}}E\Pi_{\mathcal{S}}$ is a POVM element acting only on the subspace $\mathcal{S}$. Moreover, if a POVM element is fully supported on the subspace orthogonal to $\mathcal{S}$, the probability of seeing that outcome will be zero. Thus, it suffices to consider POVMs fully supported on $\mathcal{S}$. This is a space with *complex* dimension 2, and is thus isomorphic to $\mathbb{C}^2$.

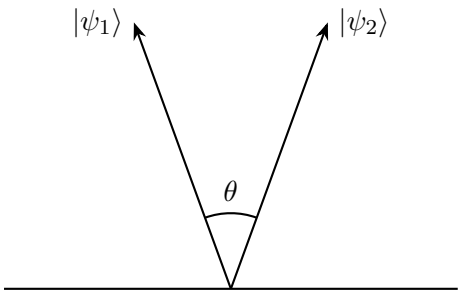

Fig. 2.1: Representation of the two vectors $|\psi_1\rangle$ and $|\psi_2\rangle$. Note that it is without loss of generality to assume $\theta \in [0, \pi/2)$. If this were not the case, one could just replace $|\psi_2\rangle$ with $-|\psi_2\rangle$. States differing by a phase factor are physically indistinguishable, so this would not change our analysis.

Okay, we have now simplified the problem considerably: we want to find a measurement in the 2-dimensional subspace $\mathcal{S}$ that minimizes the worst case error. Although we are allowed general POVMs, let us consider the simplest subset of all POVMs: measurements in a fixed basis. If we measure in a basis that biases either of the two states, Eve can exploit this information and always prepare the other state. This intuition suggests we should choose a basis that is symmetric about our two states.

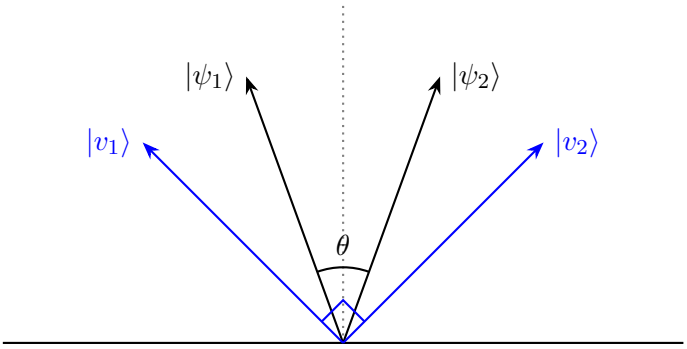

Fig. 2.2: An optimal strategy should not bias one state over the other. Among basis measurements, $\{|v_1\rangle, |v_2\rangle\}$ seems like a promising candidate.

This seems promising given the following nice features:

1. **Symmetry.** Because this basis evenly straddles the two states, our error will be symmetric! Thus, Eve cannot adversarially prepare one state over the other.

2. **Limiting cases.** When $\theta = \pi/2$, this strategy is optimal: just measure in the $\{|\psi_1\rangle, |\psi_2\rangle\}$ basis. When $\theta = 0$ (i.e. when the states are identical) the strategy succeeds with probability 1/2, which we know is optimal.

Moreover, if one tries to rotate the basis in either direction, one of the errors would decrease, but always at expense of the other. Thus the maximum error will increase if we rotate the above basis in the plane.

Okay, so what is the success probability of this strategy? Well, it is clear that the strategy treats the two states symmetrically, so it suffices to consider the probability given $|\psi\rangle = |\psi_1\rangle$. By inspecting the geometry in Fig. 2.2, we see that the angle between $|v_1\rangle$ and $|\psi_1\rangle$ is $(\pi/2 - \theta)/2$. Thus, we obtain

$$\begin{aligned} p_{\text{succ}}(\theta) &:= \Pr[\text{Guess } 1 | \text{State } 1], && (2.14)\\ &= \text{tr}\left[|v_1\rangle\langle v_1| |\psi_1\rangle\langle\psi_1|\right], && (2.15)\\ &= |\langle v_1|\psi_1\rangle|^2, && (2.16)\\ &= \cos^2\left(\frac{\pi/2 - \theta}{2}\right), && (2.17)\\ &= \frac{1}{2} + \frac{1}{2}\cos\left(\frac{\pi}{2} - \theta\right), && (2.18)\\ &= \frac{1}{2} + \frac{1}{2}\sin\theta, && (2.19) \end{aligned}$$

which is exactly what we conjectured to be optimal! We won't actually *prove* that this is optimal until next section; however, it does motivate an interesting question.

Quick Quiz 2.1.9. Suppose the above is, indeed, the optimal strategy among all possible POVMs. Should we be surprised that it is a simple basis measurement and not a more general POVM?

Pause and ponder this question! In the next section we will first prove that the above strategy *is* optimal, before returning to answer the above quick quiz in depth.

2.1.5 Exercises

Exercise 2.1.1 (Simulating Quantum Measurements)**.** In this problem, we will think about how to implement a desired measurement using the following three operations: i) appending ancillas, ii) applying unitaries, and iii) performing projective measurements in the standard basis.

(a) **Simulating[3] basis measurements.** Using only the allowable operations above, prove that we can simulate arbitrary basis measurements.

(b) **Simulating PVMs.** Do the same for general projective measurements.

(c) **Simulating arbitrary POVMs.** Suppose we have a 3-outcome POVM $\{E_1, E_2, E_3\}$. Consider the map defined as

$$|\psi\rangle \otimes |1\rangle \mapsto (\sqrt{E_1}\,|\psi\rangle) \otimes |1\rangle + (\sqrt{E_2}\,|\psi\rangle) \otimes |2\rangle + (\sqrt{E_3}\,|\psi\rangle) \otimes |3\rangle\,, \quad (2.20)$$

and similarly for the other basis elements. Compute the probability of observing the outcome "1" given the resultant state. Prove that the map, as defined, is unitary. *Note: this is a simpler version of a general statement known as Naimark's Theorem, which says that all POVMs can be implemented as PVMs on a larger Hilbert space.*

(d) **Bonus:** We discussed probabilistic strategies involving a coin flip. Construct a two-outcome POVM that implements this strategy. Then, show how to implement is as a projective measurement on a larger space.

Exercise 2.1.2 (Unambiguous State Discrimination)**.** Suppose you are given a pure quantum state $|\psi\rangle \in \mathbb{C}^d$, which is promised to either be $|\psi_1\rangle$ or $|\psi_2\rangle$. Given access to this state, you must guess "$|\psi_1\rangle$", "$|\psi_2\rangle$", or "don't know." Additionally, when the algorithm outputs "$|\psi_1\rangle$" or "$|\psi_2\rangle$," it must be correct.

[3] We say we can simulate a POVM $\{E_i\}_i$ if, using only the three allowed operations, we obtain the same probability distribution dictated by the Born rule: $p_i = \mathrm{tr}\,[\rho E_i]$.

We have seen that unless two quantum states are orthogonal, they cannot be discriminated perfectly (i.e. error probability will always be non-zero). Here, "discriminated perfectly" is taken to mean that (i) the distinguisher must always output a guess and (ii) it cannot ever be wrong. In the early days of quantum information theory, a very natural research question was: can we relax achieve (ii) if we relax (i)?

Should this even be possible? As with our above problem, it is productive to think about trivial strategies. Clearly, if we always answer "don't know," we will never misidentify the state; however, we will also never correctly identify the state. Still, this serves as a benchmark against which to test any non-trivial strategy. Naturally, the goal is to devise a scheme that uses the "don't know" response as infrequently as possible.

I encourage you to cook up strategies for this problem without looking at the rest of the problem. If you get stuck, the remaining parts will guide you towards the optimal strategy. Note, we will have Fig. 2.1 in mind as we go along.

Naturally, we would like to minimize how often we say "don't know". For the following strategies, compute the probability of saying "don't know."

(a) **Strategy 1:** Measure in the $\{|\psi_1\rangle, |\psi_1^\perp\rangle\}$ basis. Output "don't know" if the first outcome is observed and "2" if the second outcome is observed. What property does this strategy lack that an optimal strategy should have?

(b) **Strategy 2:** Flip a coin. If you observe heads, implement strategy 1, if you observe tails implement the same strategy but with respect to the $\{|\psi_2\rangle, |\psi_2\rangle^\perp\}$ basis.

(c) **Strategy 3:** Consider the collection of operators $\{E_1, E_2, E_{\text{dk}}\}$ with

$$E_1 = |\psi_2^\perp\rangle\langle\psi_2^\perp|, \quad E_2 = |\psi_1^\perp\rangle\langle\psi_1^\perp|, \quad \text{and} \quad E_{\text{dk}} = I - E_1 - E_2. \tag{2.21}$$

Explain why this does not form a valid POVM. Then, defining λ to be the largest eigenvalue of $E_1 + E_2$, show that

$$E_1 = \frac{1}{\lambda}|\psi_2^\perp\rangle\langle\psi_2^\perp|, \quad E_2 = \frac{1}{\lambda}|\psi_1^\perp\rangle\langle\psi_1^\perp|, \quad \text{and} \quad E_{\text{dk}} = I - E_1 - E_2 \tag{2.22}$$

form a valid POVM and compute the probability of saying "don't know."

If you are interested in this problem, the original literature on the topic is contained largely in Refs. [Iva87; Die88; Per88]. For pedagogical notes on the topic, see Lecture 1 of John Wright's course [Wri24].

2.2 Mixed State Discrimination

We could have started with mixed state discrimination and derived the pure state result as a corollary; however, I think the simplicity and visualizability of the pure state case make it a worthwhile starting point. In this section, we will derive the optimal strategy for distinguishing two mixed states. The problem can be stated as follows.

Problem 2.2.1 (Mixed State Discrimination)**.** Suppose we are given a mixed state $\rho \in \mathcal{D}(\mathbb{C}^d)$, which is promised to be either ρ_1 or ρ_2 (with equal probability). Determine which is the case.

In this case, we are given a prior on the two states, so we will consider the *average-case error* given as

$$p_{\text{err}}^{\text{avg}} = \frac{1}{2} \cdot \Pr[\text{Guess } \text{“}\rho_1\text{”} | \rho_2] + \frac{1}{2} \cdot \Pr[\text{Guess } \text{“}\rho_2\text{”} | \rho_1]. \tag{2.23}$$

A skeptical student might ask “why are we considering average-case error when we spent last lecture justifying the worst-case analysis?” This is a good question. The short answer is that the average case has a closed form solution which will allow us to derive analytical lower bounds on the sample complexity of various tasks. Let’s keep this question in mind and revisit it below.

2.2.1 Diagonal State Discrimination and Total Variation Distance

Before tackling the general case, let us consider the important special case when $[\rho_1, \rho_2] = 0$ (i.e. when they are simultaneously diagonalizable). Without any loss of generality, we can assume that the basis that diagonalizes these states is the standard one. In this case, the density matrices are simply two probability distributions over $[d] := \{1, 2, \ldots, d\}$ which can be written as

$$\rho_1 = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_d \end{pmatrix} \quad \text{and} \quad \rho_2 = \begin{pmatrix} q_1 & & \\ & \ddots & \\ & & q_d \end{pmatrix}. \tag{2.24}$$

Quick Quiz 2.2.2. Can you come up with a strategy for distinguishing these two states which minimizes the average case error?

Trivial Strategies

- Trivial strategy 1: flip a fair coin and choose based on the outcome! This succeeds (and fails) with probability $1/2$.

- Trivial strategy 2: always guess ρ_1 (or ρ_2). This also succeeds with probability $1/2$. These give us a benchmark we wish to exceed.

Non-trivial Strategy Remembering that we know the classical description of the two quantum states, a natural idea would be to measure in the standard basis to obtain outcome $i \in [d]$ and then simply choose the state according to $\max\{p_i, q_i\}$.

Algorithm 1 Optimal Strategy for Classical Discrimination

Require: Two probability distributions p, q over $[d]$, and a sample $x \in [d]$.
Ensure: A guess ("p" or "q") indicating the source of x.

1: **Construct the set** A of outcomes where p is more likely than q:

$$A \leftarrow \{i \in [d] : p_i \geq q_i\} = \{i \in [d] : p_i - q_i \geq 0\}$$

2: **Decision Rule:**
3: **if** $x \in A$ **then**
4: **return** "p"
5: **else** ▷ Since $x \notin A$, implies $p_x < q_x$
6: **return** "q"
7: **end if**

What is the success probability of this algorithm?

$$p_{\text{succ}} = \frac{1}{2}\sum_{x\in A} p_x + \frac{1}{2}\sum_{x\notin A} q_x, \tag{2.25}$$

$$= \frac{1}{2}\sum_{x\in A} p_x + \frac{1}{2}\left(1 - \sum_{x\in A} q_x\right), \tag{2.26}$$

$$= \frac{1}{2} + \frac{1}{2}\sum_{x\in A}(p_x - q_x), \tag{2.27}$$

$$= \frac{1}{2} + \frac{1}{2}\sum_{x\notin A}(q_x - p_x), \quad \text{Lemma 2.2.3} \tag{2.28}$$

$$= \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\sum_{x=1}^{d}|p_x - q_x|. \tag{2.29}$$

All that remains is to prove the following small lemma.

Lemma 2.2.3. Let $A := \{i \in [d] : p_i - q_i \geq 0\}$. Then,

$$\sum_{x\in A}(p_x - q_x) = \sum_{x\notin A}(q_x - p_x). \tag{2.30}$$

Proof. Because p and q are both probability distributions, we know $\sum_{x=1}^{d} p_x = \sum_{x=1}^{d} q_x = 1$. Thus, we may write

$$0 = \sum_{x=1}^{d} p_x - \sum_{x=1}^{d} q_x = \sum_{x=1}^{d} (p_x - q_x) = \sum_{x \in A} (p_x - q_x) + \sum_{x \notin A} (p_x - q_x), \tag{2.31}$$

which implies $\sum_{x \in A}(p_x - q_x) = \sum_{x \notin A}(q_x - p_x)$, as desired. □

In the next section we will rigorously prove that this strategy is optimal, but hopefully it feels like the natural thing to do.

Limiting cases. It is useful to check some limiting cases to get a feel for the performance of the algorithm. What are the limiting cases to check?

1. If $p = q$, the $p_{\text{succ}} = 1/2$, as expected. If the two distributions are equal and only Eve knows which one she gave us, our best bet is to just flip a fair coin and guess accordingly.

2. If p and q have disjoint support, our strategy will never lead us astray and we will have $p_{\text{succ}} = 1$. For example, suppose we have one coin that is heads on both sides and one that is tails on both sides (e.g. $p = (1, 0)$ and $q = (0, 1)$). Then, obtaining heads (or tails) immediately tells us which coin we were given.

$$\begin{aligned} p_{\text{succ}} &= \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \sum_{x=1}^{2} |p_x - q_x|, && (2.32) \\ &= \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} (|1 - 0| + |0 - 1|), && (2.33) \\ &= 1. && (2.34) \end{aligned}$$

Trying other examples, you can convince yourself that the quantity $\frac{1}{2} \sum_{x=1}^{2} |p_x - q_x|$ captures the distance between probability distributions. It plays such a fundamental role in classical learning and testing that is is given a name![4]

Definition 2.2.4 (Total Variation (TV) distance). Let $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ be two probability distributions on a countable probability space. The *total variation distance* between them is

$$d_{TV}(p, q) = \frac{1}{2} \cdot \sum_{x=1}^{d} |p_x - q_x|. \tag{2.35}$$

[4] See Exercise for the more general definition as well as proofs of several useful properties of the total variation distance.

Because the TV distance arises in the optimal success probability for distinguishing two probability distributions, we say that this gives the quantity an *operational interpretation*. Quantum information theorists love a good operational interpretation!

Importantly, the TV distance is a metric, meaning it satisfies the following properties:

1. **Non-negativity**. $d_{\mathrm{TV}}(p,q) \geq 0$, with equality iff $p = q$.

2. **Symmetry.** For any distributions p and q, $d_{\mathrm{TV}}(p,q) = d_{\mathrm{TV}}(q,p)$.

3. **Triangle Inequality.** For any distribution r, we have

$$d_{\mathrm{TV}}(p,q) \leq d_{\mathrm{TV}}(p,r) + d_{\mathrm{TV}}(r,q). \tag{2.36}$$

This distance is also related to an important norm that we will see a great deal throughout the course.

Definition 2.2.5 (Vector p-norm)**.** For $x = (x_1, \ldots, x_d) \in \mathbb{C}^d$, the *vector p-norm* is defined as

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}. \tag{2.37}$$

With this definition in place, we note that the TV distance is often written in terms of the 1-norm

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2} \cdot \|p - q\|_1. \tag{2.38}$$

These are very useful facts that we will see throughout the course. For now, let us return to our goal of understanding mixed state discrimination.

2.2.2 Mixed State Discrimination and Holevo-Helstrom Theorem

If you have taken a quantum information course before, it is possible that you will see a direct parallel between the above classical special case and the result to which we now turn. To state it formally, we need two additional definitions.

First, we define a matrix analogue of Def. 2.2.5.

Definition 2.2.6 (Schatten p-norm)**.** Let $M \in \mathbb{C}^{d\times d}$ be a matrix with singular values $\{\sigma_i\}_{i=1}^d$. For $p \in [1,\infty)$, the Schatten p-norm is defined as:

$$\|M\|_p := \left(\sum_{i=1}^{d} \sigma_i^p\right)^{1/p} \tag{2.39}$$

Corollary 2.2.7 (Trace Norm for Hermitian Matrices)**.** If M is Hermitian ($M = M^\dagger$) with eigenvalues $\{\lambda_i\}_{i=1}^d$, then $\sigma_i = |\lambda_i|$ and the norm becomes:

$$\|M\|_p = \left(\sum_{i=1}^{d} |\lambda_i|^p\right)^{1/p} \tag{2.40}$$

In the limit as $p \to \infty$, the norm is determined by the largest singular value. We define the Schatten ∞-norm as the **Spectral Norm**:

$$\|M\|_\infty := \max_i \sigma_i \tag{2.41}$$

Also important is the $p = 1$ case, which is typically referred to as the *trace norm* or *nuclear norm*. The trace norm will allow us to naturally define a quantum generalization of the total variation distance.

Definition 2.2.8 (Trace distance)**.** The *trace distance* between two matrices A, B is defined as

$$d_{\mathrm{tr}}(A,B) := \frac{1}{2}\|A - B\|_1 \tag{2.42}$$

There is much more to say about this distance, and we will do so next lecture. For now, we will prove the theorem that provides the main operational interpretation of the trace distance.

Theorem 2.2.9 (Holevo-Helstrom)**.** The maximal probability of distinguishing two quantum states ρ and σ is

$$p_{\mathrm{succ}}^{\max} = \frac{1}{2} + \frac{1}{2} d_{\mathrm{tr}}(\rho,\sigma), \tag{2.43}$$

$$= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}\|\rho - \sigma\|_1\right). \tag{2.44}$$

Proof. An algorithm for distinguishing two arbitrary mixed states will be to implement a two-outcome POVM $E = \{E_1, E_2\}$ and guess "ρ" when E_1 is observed and "σ" if E_2 is observed. Given this strategy, the success probability is

$$p_{\text{succ}} = \frac{1}{2}\text{tr}\left[E_1\rho\right] + \frac{1}{2}\text{tr}\left[E_2\sigma\right], \qquad \text{equal priors} \tag{2.45}$$

$$= \frac{1}{2}\text{tr}\left[E_1\rho\right] + \frac{1}{2}\text{tr}\left[(\mathbb{I} - E_1)\sigma\right], \qquad E_1 + E_2 = \mathbb{I} \tag{2.46}$$

$$= \frac{1}{2} + \frac{1}{2}\text{tr}\left[E_1(\rho - \sigma)\right], \tag{2.47}$$

where in the last line we used the linearity of trace.

Quick Quiz 2.2.10. Given that ρ and σ are both density matrices, what useful properties should we note about the operator $\rho - \sigma$?

Answer: The set of Hermitian matrices is closed under addition and real scalar multiplication, so $\rho - \sigma$ is Hermitian. Moreover, both ρ and σ have unit trace, so $\rho - \sigma$ is traceless.

Recall that the trace of a Hermitian operator is equal to the sum of the eigenvalues, thus $\sum_i \lambda_i = 0$, because $\rho - \sigma$ is traceless.

Using these facts, we can decompose the operator as

$$\rho - \sigma = \sum_{i=1}^{d} \lambda_i |v_i\rangle\langle v_i|, \tag{2.48}$$

$$= \sum_{i:\lambda_i \geq 0} \lambda_i |v_i\rangle\langle v_i| + \sum_{i:\lambda_i < 0} \lambda_i |v_i\rangle\langle v_i|, \tag{2.49}$$

$$:= P + N, \tag{2.50}$$

where P and N represent the positive and negative parts of the decomposition. Using this decomposition, we may write

$$p_{\text{succ}} = \frac{1}{2} + \frac{1}{2}\text{tr}\left[E_1(P + N)\right], \tag{2.51}$$

$$= \frac{1}{2} + \frac{1}{2}\text{tr}\left[E_1 P\right] + \frac{1}{2}\text{tr}\left[E_1 N\right]. \tag{2.52}$$

Now, we want an *upper bound* on the success probability, so what can we do? Observe that the last term can be expanded as

$$\text{tr}\left[E_1 N\right] = \sum_{i:\lambda_i<0} \lambda_i \text{tr}\left[E_1 |v_i\rangle\langle v_i|\right] = \sum_{i:\lambda_i<0} \lambda_i \underbrace{\langle v_i| E_1 |v_i\rangle}_{\geq 0} \leq 0, \tag{2.53}$$

which holds because E_i is a positive operator, by definition of a POVM. Thus dropping that term yields an upper bound. Furthermore, we know that $E_1 + E_2 = \mathbb{I}$, so $E_1 \leq \mathbb{I}$

and thus $\mathrm{tr}\left[E_1 P\right] \leq \mathrm{tr}\left[\mathbb{I} P\right] = \mathrm{tr}\left[P\right]$. Putting these together, and recalling that $\sum_i \lambda_i = 0$ because $\rho - \sigma$ is traceless, we may write

$$
\begin{aligned}
p_{\text{succ}} &\leq \frac{1}{2} + \frac{1}{2}\mathrm{tr}\left[P\right], && (2.54)\\
&= \frac{1}{2} + \frac{1}{2}\sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4}\cdot 0, && (2.55)\\
&= \frac{1}{2} + \frac{1}{2}\sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4}\sum_{i=1}^{d} \lambda_i, && (2.56)\\
&= \frac{1}{2} + \frac{1}{2}\sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4}\left(\sum_{i:\lambda_i \geq 0} \lambda_i + \sum_{i:\lambda_i < 0} \lambda_i\right), && (2.57)\\
&= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}\sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{2}\sum_{i:\lambda_i < 0} \lambda_i\right), && (2.58)\\
&= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}\sum_{i:\lambda_i \geq 0} |\lambda_i| + \frac{1}{2}\sum_{i:\lambda_i < 0} |\lambda_i|\right), && (2.59)\\
&= \frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\sum_{i=1}^{d} |\lambda_i|, && (2.60)\\
&= \frac{1}{2} + \frac{1}{2}\|\rho - \sigma\|_1, \quad \text{Def. 2.2.6} && (2.61)\\
&= \frac{1}{2} + \frac{1}{2} d_{\mathrm{tr}}(\rho, \sigma), && (2.62)
\end{aligned}
$$

where the last line follows from Def. 2.2.8. This is an amazingly simple, but essential result that we will use again and again throughout this course.

Quick Quiz 2.2.11. Anytime we prove an inequality in this course, it is important to ask under what conditions the inequality is saturated. So, in this case, can this inequality be saturated and, if so, under what conditions?

To derive this upper bound, we used $\mathrm{tr}\left[E_1 N\right] \leq 0$ and $\mathrm{tr}\left[E_1 P\right] \leq \mathrm{tr}\left[P\right]$. Thus, to achieve equality, we need E_1 to have no overlap with N, and maximal overlap with P. If we set $E_1 = \sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i|$, we have

$$
\begin{aligned}
\mathrm{tr}\left[E_1 N\right] &= \mathrm{tr}\left[\sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i| \sum_{j:\lambda_j < 0} \lambda_j |v_j\rangle\langle v_j|\right], && (2.63)\\
&= 0, \quad \text{orthonormality of } \{|v_i\rangle\} && (2.64)
\end{aligned}
$$

as well as

$$\begin{aligned}
\operatorname{tr}\left[E_1 P\right] &= \operatorname{tr}\left[\sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i| \sum_{i:\lambda_i \geq 0} \lambda_i |v_j\rangle\langle v_j|\right], && (2.65)\\
&= \operatorname{tr}\left[\sum_{i:\lambda_i \geq 0} \lambda_i |v_j\rangle\langle v_j|\right], && (2.66)\\
&= \sum_{i:\lambda_i \geq 0} \lambda_i, && (2.67)\\
&= \operatorname{tr}\left[P\right], && (2.68)
\end{aligned}$$

as desired. Thus, we have shown that the optimal POVM is defined by $E = \{E_1, E_2\}$ with E_1 spanned by the eigenvectors of $\rho - \sigma$

$$E_1 = \sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i| \quad \text{and} \quad E_2 = I - E_1, \tag{2.69}$$

which yields a maximum probability of success given as

$$p_{\text{succ}}^{\max} = \frac{1}{2} + \frac{1}{2} d_{\text{tr}}(\rho, \sigma). \tag{2.70}$$

□

There is a less instructive, but streamlined proof of this result using Hölder's inequality for Hermitian matrices (see Exercise 2.2.3). I also encourage you to attempt Exercises so that you see formally how to derive the special cases we considered from Holevo-Helstrom.

2.2.3 Exercises

Exercise 2.2.1 (Optimal pure state distinguishing)**.** Use Theorem 2.2.9 to prove our optimal pure state distinguishing formula

$$p_{\text{succ}}(\theta) = \frac{1}{2} + \frac{1}{2} \sin\theta. \tag{2.71}$$

Exercise 2.2.2 (Hölder's Inequality for Hermitian Matrices)**.** Given two Hermitian matrices A, B and $p, q \in [1, \infty)$ such that $\frac{1}{p} + \frac{1}{q} = 1$, prove that

$$\operatorname{tr}\left[AB\right] \leq \|A\|_p \|B\|_q. \tag{2.72}$$

Hint: for an excellent treatment of Hölder's inequality for real numbers (which is needed to prove this result), as well as many other wonderful inequalities, see Ref. [Ste04].

Exercise 2.2.3 (Alternate Proof of Holevo-Helstrom)**.** Using Eq. (2.72), provide an alternate proof of Theorem 2.2.9.

2.3 Discrimination with Multiple Samples

In the previous section, we studied the problem of discriminating two unknown states or distributions given only one sample. This culminated with Theorem 2.2.9, which gives operational meaning to the trace distance and, as such, will play a fundamental role in proving sample complexity results in this course.

Sadly, if the states (or distributions) are very close to one another, we won't be able to do much better than randomly guessing. Suppose, however, that we can pay Eve for additional samples.

> **Quick Quiz 2.3.1.** Will having access to more samples from either p or q help us distinguish these samples?

The typical student response is: of course! But when pressed to provide a precise mathematical justification, they are less certain. Let us now formalize this intuition.

The set-up is now as follows. Eve will select one of two machines with equal probability. Every time we press the button, we pay a dollar for a new sample of either p or q. Only Eve knows which is the case, and our goal is to determine (with high-probability) which distribution we are sampling from using as few samples as possible.

Suppose Eve selected p. Then, for all $i \in [n]$, $x_i \sim p$. After pressing the button n times, we have n samples which we can collect in a vector

$$\boldsymbol{x} := (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n. \tag{2.73}$$

Because we have assumed that the samples are independent and identically distributed (i.i.d), we will denote the distribution over all 2^n possible $\boldsymbol{x}$'s as

$$p^{\otimes n} := p_{x_1} p_{x_2} \cdots p_{x_n}. \tag{2.74}$$

If you haven't seen this notation before, note that it originates naturally when representing probability distributions as random vectors. For example, consider two independent coin flips. If $p(\text{heads}) = a$ and $p(\text{tails}) = b$, we can write

$$p = \begin{bmatrix} a \\ b \end{bmatrix} \implies p^{\otimes 2} = \begin{bmatrix} a^2 \\ ab \\ ba \\ b^2 \end{bmatrix}. \tag{2.75}$$

This generalizes naturally to n independent samples. In this setting, then, the goal becomes distinguishing between $p^{\otimes n}$ and $q^{\otimes n}$. Let's consider a concrete example that will guide our intuition.

Quick Quiz 2.3.2 (Fair vs Biased Coin). Let $\epsilon \in (0, 1)$. Suppose you are given n samples of either $p = (1/2, 1/2)$ or $q = (1/2 + \epsilon, 1/2 - \epsilon)$. How many samples suffice to distinguish these two cases with probability at least 0.99?

Insert binary tree representation and histograms.

Well, we saw in the last section that when quantum states are simultaneously diagonalizable, Holevo-Helstrom (Theorem 2.2.9) upper bounds the success probability of distinguishing distributions. Thus, any algorithm used to distinguish $p^{\otimes n}$ from $q^{\otimes n}$ must satisfy

$$p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2} d_{\text{TV}}(p^{\otimes n}, q^{\otimes n}). \tag{2.76}$$

If we want to succeed with probability at least 0.99, then we need

$$0.99 \leq p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2} d_{\text{TV}}(p^{\otimes n}, q^{\otimes n}) \implies 0.98 \leq d_{\text{TV}}(p^{\otimes n}, q^{\otimes n}). \tag{2.77}$$

Given p and q, is this easy to compute? Well, for general discrete distributions over $[d]$, there are d^n terms in the sum needed to compute the TV distance, so the classical computation cost will scale exponentially in the number of samples. We will return to this point at the end of the lecture, but for now, lets cook up an algorithm that would allow us to distinguish between the two cases.

2.3.1 Distinguishing Probability Distributions with Multiple Samples

Quick Quiz 2.3.3. Can you come up with a simple algorithm to distinguish between a fair and biased coin, given n samples?

To formalize things, let us recall the definition of a Bernoulli random variable.

Definition 2.3.4 (Bernoulli Random Variable)**.** A random variable X is said to have a *Bernoulli distribution* with probability of success $\alpha \in [0, 1]$, denoted as $X \sim \text{Bern}(\alpha)$, if its probability mass function (PMF) is

$$P(X = x) = \begin{cases} \alpha & \text{if } x = 1 \\ 1 - \alpha & \text{if } x = 0. \end{cases} \tag{2.78}$$

Recall, also, that

$$\mathbb{E}[X] = \alpha \cdot 1 + (1 - \alpha) \cdot 0 = \alpha, \tag{2.79}$$

$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \alpha - \alpha^2 = \alpha(1 - \alpha) \tag{2.80}$$

Given n coin flips, then, we have $\boldsymbol{x} \sim p^{\otimes n}$ which are a distribution over *ordered* length-n bit strings. Hopefully you can convince yourself that order should not matter when attempting to distinguish two Bernoulli random variables. Perhaps the most natural algorithm is to simply count the number of heads that appear in our sequence of n outcomes. Let's define the random variable

$$K = \sum_{i=1}^{n} X_i, \tag{2.81}$$

where $X_i \sim \text{Bern}(\alpha)\ \forall\ i \in [n]$. If we imagine the leaf nodes of our binary tree above, the probability that a particular leaf has k heads is given as $\alpha^k(1 - \alpha)^{n-k}$. But, as mentioned above, order does not matter here, so the actual probability of obtaining k heads is simply this probability times the number of ways to have a length n bit string with k ones (heads)

$$p(K = k) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} =: \text{Bin}(n, \alpha), \tag{2.82}$$

which is the PMF of a Binomial random variable. Due to the independence of the trials, we can easily compute

$$\mathbb{E}[K] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\alpha, \tag{2.83}$$

$$\text{Var}[K] = \sum_{i=1}^{n} \text{Var}[X_i] = n\alpha(1 - \alpha). \tag{2.84}$$

We may now formalize the intuition that more samples will help us distinguish between a fair and biased coin. Recall that the central limit theorem says that a

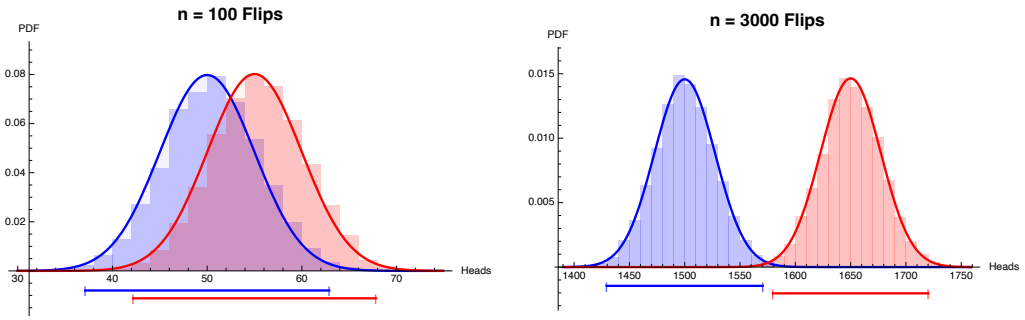

Fig. 2.3: Two histograms showing the distribution of the number of heads given n flips, with the 99% confidence intervals superposed.

Binomial distribution will approach a Gaussian (normal) distribution[5] in the limit of large n.

Figure 2.3 indicates that taking more samples will make the distributions more distinguishable, but how many samples suffice? To determine this, we need our first *concentration inequality*.

Theorem 2.3.5 (Chebyshev's Inequality)**.** If X is a real-valued random variable, then for any $c > 0$, we have

$$\Pr[|X - \mathbb{E}[X]| \geq c \cdot \sqrt{\mathrm{Var}[X]}] \leq \frac{1}{c^2}, \tag{2.85}$$

or, equivalently,

$$\Pr[|X - \mathbb{E}[X]| \geq c] \leq \frac{\mathrm{Var}[X]}{c^2}. \tag{2.86}$$

We note that, while it will be straightforward in our case, we do not actually need to exactly compute the variance of a random variable to apply Chebyshev, we only need an upper bound. Later in this course, we will see examples for which it is tedious or intractable to compute the variance exactly, so it is worth noting that a good upper bound suffices.

For our problem, we need to compute or bound the variance of a random variable $K \sim \mathrm{Bin}(n, \alpha)$. We may write

$$\mathrm{Var}[K] = n\alpha(1-\alpha) \leq \frac{n}{4}, \tag{2.87}$$

where the inequality comes from observing that the maximum of $\alpha - \alpha^2$ occurs when $\alpha = \frac{1}{2}$. Now, we want to succeed with probability at least 0.99, so we must take n

[5] If you have not seen this demonstrated with a Galton board, check this video out!

sufficiently large to ensure that the 99% condifence intervals shown in Fig. 2.3 do not overlap. That is, we to find n such that

$$\Pr\left[\left|K - \frac{n}{2}\right| \geq 10\sqrt{n/4}\right] \leq \frac{1}{100}. \tag{2.88}$$

To distinguish the $\mathrm{Bin}(n, \frac{1}{2})$ from $\mathrm{Bin}(n, \frac{1}{2}+\epsilon)$, we can find the midpoint between the two means and ensure that our 99% confidence intervals meet there. The midpoint is given as

$$\frac{1}{2}\cdot\frac{1}{2}n + \frac{1}{2}\cdot\left(\frac{1}{2}+\epsilon\right)n = \left(\frac{1}{2}+\frac{1}{2}\epsilon\right)n. \tag{2.89}$$

It follows that the distance between the fair mean and the midpoint is

$$\left(\frac{1}{2}+\frac{1}{2}\epsilon\right)n - \frac{1}{2}\cdot n = \frac{\epsilon}{2}. \tag{2.90}$$

Thus, we need to choose n such

$$\Pr\left[\left|K - \frac{n}{2}\right| \geq \frac{\epsilon}{2}\right] \leq \Pr\left[\left|K - \frac{n}{2}\right| \geq 5\sqrt{n}\right] \leq \frac{1}{100}. \tag{2.91}$$

The first inequality holds only when $n\epsilon/2 \geq 5\sqrt{n}$ or, equivalently, when

$$n \geq \frac{100}{\epsilon^2} \implies n = O\left(\frac{1}{\epsilon^2}\right). \tag{2.92}$$

This is an upper bound on the *sample complexity* of distinguishing a fair coin from a slightly biased one. The so-called big-O notation[6], $O(\cdot)$, essentially hides any constants and lets one focus on the asymptotic scaling with the parameter of interest. We will develop more advanced tools as we proceed through the course, but the general strategy for deriving sample complexity upper bounds will always involve some sort of concentration inequality.

When we prove a sample complexity upper bound, we should always ask if the result is tight. That is: do we always *need* this many samples to solve the problem?

> **Quick Quiz 2.3.6.** Can you think of two distributions that should take fewer samples to distinguish?

Consider, for example, the distributions

$$p' = (1, 0), \tag{2.93}$$

$$q' = (1-\epsilon, \epsilon). \tag{2.94}$$

[6] If you have not ever seen Big-O notation, please go watch Ryan O'Donnell's lecture on the topic before continuing!

Thinking of these as coins again, we should be able to simply flip the coin $n = O\left(\frac{1}{\epsilon}\right)$ times and be reasonably confident which case we are in because, in the first case, we are guaranteed not to see tails. Thus, if we see even one tails, we know we are sampling from q'. In Exercise 2.3.1, you will formalize this. This highlights a shortcoming of the total variation distance. Notice that

$$d_{\mathrm{TV}}(p, q) = \epsilon, \tag{2.95}$$

$$d_{\mathrm{TV}}(p', q') = \epsilon, \tag{2.96}$$

and yet they have drastically different sample complexity upper bounds. In the next section, we will meet a distance measure that more satisfactorily captures the difference between these two cases and allows us to determine the sample complexity of distinguishing two distributions easily.

2.3.2 Exercises

Exercise 2.3.1. Let $p' = (1, 0)$, $q' = (1 - \epsilon, \epsilon)$, and $\delta \in (0, 1)$. Show that there exists an algorithm using $n = O\left(\frac{\log 1/\delta}{\epsilon}\right)$ samples to distinguish between $p'^{\otimes n}$ and $q'^{\otimes n}$ with probability at least $1 - \delta$.

3 Quantum State Tomography

> *Quantum state tomography['s] perfection is of great importance to quantum computation and quantum information.*
>
> **— Nielsen and Chuang**

Quantum state tomography (QST) is the task of learning a classical description of an unknown quantum state. If you accept that the density matrix is the complete description of the underlying quantum system, then it follows that any property of that quantum state should be able to be approximated by appropriately post-processing this classical approximation of the quantum state. In this sense, QST is perhaps the most fundamental task in quantum learning theory.

The first mention of quantum state tomography in the literature, as far as I can tell, is in a 1987 paper entitled "A tomographic approach to Wigner's function" by Bertrand and Bertrand [BB87]. The name seems to have originated due to the analogy with computerized tomography (CT) scans in the medical field. The first experiment actually implementing such a tomographic procedure was due to Smithey *et al.* [Smi+93]. Many of the early references to tomography were actually in the continuous variable setting, because quantum optics experiments reached sufficient maturity before finite-dimensional quantum systems. The theoretical extension to finite-dimensional systems followed shortly after in Ref. [Leo95] (which is also the first paper, to my knowledge, that uses the term "quantum state tomography").

With this history in mind, let us define the task formally.

Definition 3.0.1 (Quantum State Tomography)**.** Let $\epsilon, \delta \in (0, 1)$. Given n copies of a quantum state ρ, output a classical description of the state, $\hat{\rho}$, such that

$$\Pr\left[d_{\mathrm{tr}}(\rho, \hat{\rho}) \leq \epsilon\right] > 1 - \delta. \tag{3.1}$$

As far as quantum learning theory is concerned, we would like to determine how many copies (or samples) n are necessary and sufficient for this task. This value will change depending on what distance measure is used and what measurements

are allowed. In this course, we will primarily focus on tomography with respect to trace distance, due to its operational meaning (c.f. Theorem 2.2.9), though we know that tomography with respect to other distance measures has been considered in the literature.

This chapter will essentially be a detailed look at how the resources we have access to change the sample complexity of QST. Because QST is a canonical example of a quantum learning task, studying it closely will pay dividends in the rest of the course. Before we get started, and so we are all on the same page, let us take a moment to define the main measurement classes we will consider.

3.1 Measurement Classes

We will add more granularity, figures, and discussion later.

Definition 3.1.1 (Locality of POVMs)**.** Let $\mathcal{H} = \left((\mathbb{C}^d)^{\tilde{\otimes} n}\right)^{\otimes T}$ denote the Hilbert space for T copies of an n-qudit quantum system, where the "$\tilde{\otimes}$" distinguishes the tensor product *within* copies from the tensor product *between* copies. We consider three levels of measurement locality, ordered from most to least powerful:

- **Multi-copy:** an arbitrary POVM on $\mathcal{H}$, with no restriction on entanglement across copies or qudits.
- **Single-copy, global:** a POVM whose elements factor across the T copies, i.e., each measurement is an unrestricted POVM on a single copy $(\mathbb{C}^d)^{\tilde{\otimes} n}$, applied independently to each copy.
- **Single-copy, local:** a POVM whose elements factor across both the T copies and the n qudits within each copy, i.e., each measurement is a product of single-qudit POVMs on $\mathbb{C}^d$.

3.2 Single-copy, Local Tomography Algorithms

While single-copy, local measurements are the most restrictive (and thus the least informative), they are also the most experimentally-friendly. In fact, they remain the standard in many experimental labs. Moreover, there are very simple algorithms in this class of measurements, so it is a good starting point conceptually, as it will anchor our future approaches.

3.2.1 Pauli Matrices Crash Course

There is much to say about the Pauli matrices. For now, we will just review the essential properties needed to understand the standard Pauli tomography algorithm.

We will denote the set of all n-qubit Pauli matrices as

$$\mathcal{P}_n = \{P = P_1 \otimes \cdots \otimes P_n | P_i \in \{I, X, Y, Z\}\}. \tag{3.2}$$

The elements of this set are often referred to as “Pauli strings" due to a correspondence with Boolean algebra that can be made rigorous (more on this down the line). For now, let us just state the most relevant results for the n-qubit Paulis.

Proposition 3.2.1 (Properties of n-qubit Paulis)**.** All n-qubit Pauli matrices satify the following properties:

1. **Hermiticity.** $P = P^\dagger$ with eigenvalues ± 1.
2. **Involutory.** $P^2 = \mathbb{I}$.
3. **Traceless.** $\mathrm{tr}\,[P] = 0$ for all $P \in \mathcal{P}_n \setminus \{\mathbb{I}\}$
4. **Orthogonality.** If $P_i, P_j \in \mathcal{P}_n$, then $\mathrm{tr}\,[P_i P_j] = \delta_{ij} 2^n$.

From these properties, one can prove the following lemma.

Lemma 3.2.2 (Pauli Bases)**.** The n-qubit Pauli matrices form basis for the following finite-dimensional vector spaces:

1. the 4^n-dimensional complex vector space $\mathbb{C}^{2^n \times 2^n}$ of $2^n \times 2^n$ complex matrices,
2. the 4^n-dimensional real vector space of $2^n \times 2^n$ Hermitian matrices.

Both will be useful at times; however, the latter is the most important for us at present because quantum states are a subset of all $2^n \times 2^n$ Hermitian matrices. The “textbook" Pauli tomography algorithm utilizes this fact, as we will now see.

3.2.2 Textbook Pauli Tomography

Tomography using binary Pauli measurements is perhaps the most straight-forward tomography algorithms conceivable, and is still the workhorse of many small-scale experimental efforts. That said, there was not (to my knowledge) a standard reference deriving the sample complexity of the algorithm in the literature. So, let us do so now.

In light of Lemma 3.2.2, we know that any n-qubit quantum state, $\rho \in (\mathbb{C}^2)^{\otimes n}$, can be expressed as

$$\rho = \frac{1}{2^n} \sum_{i=1}^{d^2} \mathrm{tr}\left[P_i \rho\right] P_i, \tag{3.3}$$

where the $P_i \in \{\mathbb{I}, \sigma_x, \sigma_y, \sigma_z\}^{\otimes n}$. The algorithm is simple: measure each coefficient, $\alpha_{P_i} := \mathrm{tr}\left[P_i \rho\right]$ with accuracy sufficient to guarantee our overall estimated state will be close to the actual state in some desired distance measure. That is, for all $i \in [d^2]$, we implement the projective measurement[1] $\{\frac{1}{2}(\mathbb{I} + P_i), \frac{1}{2}(\mathbb{I} - P_i)\}$, with possible outcomes $x_i \in \{\pm 1\}$.

Algorithm 2 "Textbook" Pauli Tomography

Require: n-qubit state ρ, number of samples per Pauli string M.

1: **for** each non-identity $P_i \in \mathcal{P}_n \setminus \{I\}$ **do**
2: Measure the P_i observable M times. Receive $x_i^{(m)} \in \{\pm 1\}$ for all $m \in [M]$.
3: $\hat{\alpha}_{P_i} \leftarrow \frac{1}{M} \sum_{m=1}^{M} x_i^{(m)}$
4: **end for**
5: Set $\hat{\alpha}_I \equiv 1$
6: **return** $\hat{\rho} = \frac{1}{2^n} \sum_{i=1}^{d^2} \hat{\alpha}_{P_i} \cdot P_i$

Suppose we prepare ρ and carry out the i^{th} Pauli measurement M times. The natural estimator of the i^{th} Pauli coefficient is simply the empirical average of these measurement outcomes

$$\hat{\alpha}_{P_i} := \frac{1}{M} \sum_{m=1}^{M} x_i^{(m)}. \tag{3.4}$$

[1] This looks like a single-copy, global measurement. Think about how you might simulate it with single-copy, local measurements and classical post-processing.

To see that this estimator is unbiased, observe

$$\mathbb{E}\left[\hat{\alpha}_{P_i}\right] = \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} x_i^{(m)}\right], \tag{3.5}$$

$$= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[x_i^{(m)}\right], \tag{3.6}$$

$$= \mathbb{E}\left[x_i^{(m)}\right], \tag{3.7}$$

$$= (+1)\cdot \Pr[+1] + (-1)\cdot \Pr[-1], \tag{3.8}$$

$$= \operatorname{tr}\left[\frac{1}{2}(\mathbb{I}+P_i)\rho\right] - \operatorname{tr}\left[\frac{1}{2}(\mathbb{I}-P_i)\rho\right], \tag{3.9}$$

$$\mathbb{E}\left[\hat{\alpha}_{P_i}\right] = \operatorname{tr}\left[P_i\rho\right]. \tag{3.10}$$

By linearlity of the expectation, it follows immediately that $\hat{\rho} = \frac{1}{2^n}\sum_{i=1}^{4^n} \hat{\alpha}_{P_i} P_i$ is an unbiased estimator of ρ (i.e. $\mathbb{E}\left[\hat{\rho}\right] = \rho$). Before we proceed, let us state an often useful lemma.

Lemma 3.2.3 (Equivalence of Schatten 1 and 2 Norms)**.** For any matrix $A \in \mathbb{C}^{d\times d}$, the Schatten 1-norm (trace norm) $\|A\|_1$ and the Schatten 2-norm (Frobenius norm) $\|A\|_2$ satisfy the following inequalities:

$$\|A\|_2 \leq \|A\|_1 \leq \sqrt{d}\|A\|_2 \tag{3.11}$$

These inequalities will be used frequently, so take the time to prove them! As a hint, the lower bound follows from directly comparing the square of both norms and the upper bound follows from everyone's favorite inequality[2].

This lemma is useful because it is much easier to work with the 2-norm and then convert to the 1-norm in the end. Moreover, we eventually want a statement about the closeness of the approximation *with high probability*; however, it is also easier to bound things in expectation and then convert to a statement in probability at the end.

Quick Quiz 3.2.4. I claim that to obtain close approximation in trace distance, and with high probability, it suffices to upper bound

$$\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right]. \tag{3.12}$$

Is this true? Is there an intuitive reason?

[2] Cauchy-Schwarz... always try Cauchy-Schwarz.

First, the intuition can be obtained from expanding the expression:

$$\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right] = \mathbb{E}\left[\frac{1}{2^n}\sum_{P\in\mathcal{P}_n}(\alpha_P - \hat{\alpha}_P)^2\right], \tag{3.13}$$

$$= \frac{1}{2^n}\sum_{P\in\mathcal{P}_n}\mathbb{E}\left[(\alpha_P - \hat{\alpha}_P)^2\right], \tag{3.14}$$

$$= \frac{1}{2^n}\sum_{P\in\mathcal{P}_n}\mathrm{Var}\left[\hat{\alpha}_P\right]. \tag{3.15}$$

Thus, the expected value of the squared Euclidean distance between the actual state and our estimate is proportional to the sum of variances of each coefficient estimator. Chebyshev's inequality should then give the intuition that upper bounding this variance should concentrate the values around their means (i.e. the true value).

We can formalize this intuition as follows. We want to show that an upper bound on $\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right]$ implies an upper bound on $\|\rho - \hat{\rho}\|_1$ with high probability. Observe

$$\mathbb{E}\left[\|\rho - \hat{\rho}\|_1\right] \leq \sqrt{\mathbb{E}\left[\|\rho - \hat{\rho}\|_1^2\right]}, \qquad \mathrm{Var}\left[X\right] \geq 0, \tag{3.16}$$

$$\leq \sqrt{d\cdot\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right]}, \qquad \text{Lemma 3.2.3}, \tag{3.17}$$

$$= \sqrt{d}\sqrt{\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right]}. \tag{3.18}$$

To convert this to a bound that holds "with high probability," we will need Markov's inequality.

Theorem 3.2.5 (Markov's Inequality)**.** Let X be a non-negative random variable and $a > 0$. Then

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{3.19}$$

Equivalently, setting $a = c\cdot\mathbb{E}[X]$ for $c > 0$,

$$\Pr[X \geq c\cdot\mathbb{E}[X]] \leq \frac{1}{c}. \tag{3.20}$$

We can apply this to the random variable $\|\rho - \hat{\rho}\|_2^2$ as follows. Suppose we prove

$$\mathbb{E}\left[\|\rho - \hat{\rho}\|_2^2\right] \leq \frac{d}{M}. \tag{3.21}$$

Then, the second form of Markov's implies

$$\Pr[\|\rho - \hat{\rho}\|_2^2 \geq c\cdot\frac{d}{M}] \leq \frac{1}{c}. \tag{3.22}$$

Setting $c = 100$, for example, yields

$$\Pr\left[\|\rho - \hat{\rho}\|_2^2 \geq 100 \cdot \frac{d}{M}\right] \leq \frac{1}{100} \quad \iff \quad \|\rho - \hat{\rho}\|_2 < O\left(\sqrt{\frac{d}{M}}\right), \quad \text{w.h.p.} \tag{3.23}$$

Then, a simple application of Lemma 3.2.3 yields

$$\|\rho - \hat{\rho}\|_1 \leq \sqrt{d}\|\rho - \hat{\rho}\|_2 < O\left(\sqrt{\frac{d^2}{M}}\right), \quad \text{w.h.p.} \tag{3.24}$$

We want an ϵ-close approximation, so we can set $\epsilon := O\left(\sqrt{d^2/M}\right)$. We must then take $M = O(d^2/\epsilon^2)$ samples per Pauli to achieve this. There are d^2 total Paulis, so the total sample complexity becomes

$$T = d^2 \cdot M = O\left(\frac{d^4}{\epsilon^2}\right) = O\left(\frac{16^n}{\epsilon^2}\right). \tag{3.25}$$

We will have our sample complexity upper bound if we can prove $\mathbb{E}\left[\|\hat{\rho} - \rho\|_2^2\right] \leq d/M$. Let us now show this.

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\rho} - \rho\|_2^2\right] &= \mathbb{E}\left[\operatorname{tr}\left[(\hat{\rho} - \rho)^\dagger(\hat{\rho} - \rho)\right]\right], && (3.26)\\
&= \mathbb{E}\left[\frac{1}{d^2}\sum_{i=1}^{d^2}\sum_{j=1}^{d^2}(\hat{\alpha}_{P_i} - \operatorname{tr}\left[P_i\rho\right])(\hat{\alpha}_{P_j} - \operatorname{tr}\left[P_j\rho\right]) \underbrace{\operatorname{tr}\left[P_iP_j\right]}_{d\delta_{ij}}\right], && (3.27)\\
&= \mathbb{E}\left[\frac{1}{d}\sum_{i=1}^{d^2}(\hat{\alpha}_{P_i} - \operatorname{tr}\left[P_i\rho\right])^2\right], && (3.28)\\
&= \frac{1}{d}\sum_{i=1}^{d^2}\operatorname{Var}\left[\hat{\alpha}_{P_i}\right], && (3.29)\\
&= \frac{1}{d}\sum_{i=1}^{d^2}\operatorname{Var}\left[\frac{1}{M}\sum_{m=1}^{M}x_i^{(m)}\right], && (3.30)\\
&= \frac{1}{dM^2}\sum_{i=1}^{d^2}\sum_{m=1}^{M}\operatorname{Var}\left[x_i^{(m)}\right], && (3.31)\\
&= \frac{1}{dM^2}\sum_{i=1}^{d^2}\sum_{m=1}^{M}\left(\mathbb{E}\left[(x_i^{(m)})^2\right] - \mathbb{E}\left[x_i^{(m)}\right]^2\right), && (3.32)\\
&= \frac{1}{dM^2}\sum_{i=1}^{d^2}\sum_{m=1}^{M}\left(1 - \mathbb{E}\left[x_i^{(m)}\right]^2\right), && (3.33)\\
&\leq \frac{1}{dM^2}\sum_{i=1}^{d^2}\sum_{m=1}^{M}1, && (3.34)\\
\implies \mathbb{E}\left[\|\hat{\rho} - \rho\|_2^2\right] &\leq \frac{d}{M}. && (3.35)
\end{aligned}$$

3.2.3 Additional Notes on Textbook Pauli Tomography

Most references that describe the Pauli tomography algorithm, do so roughly as we have done above. An attentive student asked "how is this a single-copy, local strategy?" They rightly pointed out that $\{\frac{1}{2}(\mathbb{I}+P_i), \frac{1}{2}(\mathbb{I}-P_i)\}$ is not the same as the product of local projectors onto individual Paulis. This is absolutely true. However, the algorithm we presented only utilized the eigenvalue obtained (the post-measurement state is irrelevant). In such a case, the single-copy, global measurement statistics can be simulated by single-copy, local measurements with post-processing.

Lemma 3.2.6 (Local Simulation of Global Pauli Projections)**.** Suppose $P \in \mathcal{P}_n$ is an n-qubit Pauli string and the corresponding two-outcome PVM $\{\frac{1}{2}(\mathbb{I}+P), \frac{1}{2}(\mathbb{I}-P)\}$ which yields outcome $b \in \{+1, -1\}$ with probability

$$\Pr_{\text{global}}[b] = \text{tr}\left[\rho\left(\frac{1+bP}{2}\right)\right] = \frac{1+b\text{tr}\left[\rho P\right]}{2}. \tag{3.36}$$

Moreover, consider the local protocol:

1. For each qubit $i \in [n]$, measure $\{\frac{1}{2}(I+P_i), \frac{1}{2}(I-P_i)\}$ obtaining $b_i \in \{+1, -1\}$
2. Output $b = \prod_{i=1}^{n} .$

Then, for any state ρ and outcome $b \in \{+1, -1\}$, we have

$$\Pr_{\text{global}}[b] = \Pr_{\text{local}}[b]. \tag{3.37}$$

Proof. Because the local Pauli measurements are independent, the joint probability of obtaining $b_i \in \{+1, -1\}$ on each qubit $i \in [n]$ is

$$\Pr_{\text{local}}[b_1, \ldots, b_n] = \text{tr}\left[\rho \bigotimes_{i=1}^{n} \frac{I_i + b_i P_i}{2}\right]. \tag{3.38}$$

Then, the probability of obtaining a string satisfying $b = \prod_{i=1}^{n} b_i$ is the sum over all such strings

$$\Pr_{\text{local}}[b] = \sum_{\substack{b_1,\ldots,b_n \in \{+1,-1\} \\ \prod_i b_i = b}} \operatorname{tr}\left[\rho \bigotimes_{i=1}^{n} \frac{I_i + b_i P_i}{2}\right], \tag{3.39}$$

$$= \sum_{b_1,\ldots,b_n \in \{+1,-1\}} \mathbf{1}\left[\prod_j b_j = b\right] \operatorname{tr}\left[\rho \bigotimes_{i=1}^{n} \frac{I_i + b_i P_i}{2}\right], \tag{3.40}$$

$$= \sum_{b_1,\ldots,b_n \in \{+1,-1\}} \frac{1}{2}\left(1 + b \prod_j b_j\right) \operatorname{tr}\left[\rho \bigotimes_{i=1}^{n} \frac{I_i + b_i P_i}{2}\right], \tag{3.41}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, which we have simplified using the fact that for $x, y \in \{\pm 1\}$,

$$\frac{1 + bc}{2} = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{else.} \end{cases} \tag{3.42}$$

Now, we may expand the tensor product much as we would if we were just using the standard binomial theorem. One can show that

$$\bigotimes_{i=1}^{n} \frac{I_i + b_i P_i}{2} = \frac{1}{2^n} \sum_{S \subseteq [n]} \left(\prod_i b_i\right) \bigotimes_{i=1}^{n} P_i^{(S)}, \tag{3.43}$$

where $P_i^{(S)} = P_i$ if $i \in S$ and $P_i^{(S)} = I_i$ if $i \notin S$. Plugging this back into our expression, and using linearity of the trace, we obtain

$$\Pr_{\text{local}}[b] = \frac{1}{2} \cdot \frac{1}{2^n} \sum_{S \subseteq [n]} \operatorname{tr}\left[\rho \bigotimes_{i=1}^{n} P_i^{(S)}\right] (\Sigma_1(S) + b \cdot \Sigma_2(S)), \tag{3.44}$$

where we have defined

$$\Sigma_1(S) := \sum_{b_1,\ldots,b_n \in \{+1,-1\}} \prod_{i \in S} b_i, \tag{3.45}$$

$$\Sigma_2(S) := \sum_{b_1,\ldots,b_n \in \{+1,-1\}} \prod_{i \in S} b_i \cdot \prod_{j=1}^{n} b_j. \tag{3.46}$$

Then, noting that $\prod_{i \in \emptyset} b_i := 1$, one can show

$$\Sigma_1(S) = \begin{cases} 2^n, & S = \emptyset, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \Sigma_2(S) = \begin{cases} 2^n, & S = [n], \\ 0, & \text{otherwise.} \end{cases} \tag{3.47}$$

Plugging this back into our expression, we have

$$\Pr_{\text{local}}[b] = \frac{1}{2} \cdot \frac{1}{2^n} \sum_{S \subseteq [n]} \operatorname{tr}\left[\rho \bigotimes_{i=1}^{n} P_i^{(S)}\right] (\Sigma_1(S) + b \cdot \Sigma_2(S)), \tag{3.48}$$

$$= \frac{1}{2} \cdot \frac{1}{2^n} (2^n \cdot \operatorname{tr}[\rho] + 2^n \cdot b \cdot \operatorname{tr}[\rho P]), \tag{3.49}$$

$$= \frac{1 + b \cdot \operatorname{tr}[\rho P]}{2}, \tag{3.50}$$

$$= \Pr_{\text{global}}[b], \tag{3.51}$$

as desired. □

This is possible because the single-copy, global measurement is essentially a *coarse-graining* of the single-copy, local measurement outcomes. In other words, in each single-copy, local measurement round, we obtain n bits of information, which we can post-process to give the 1 bit of information that is obtained from the global measurement. While this confirms our intuition that the above algorithm can be considered single-copy, local, it also highlights an inefficiency: we are throwing away information that might be useful!

It turns out that if you restrict to non-adaptive, single-copy measurements with $O(1)$ outcomes, then the “textbook" Pauli tomography algorithm is actually optimal (see Corollary 4.7 in Ref. [LN25] as well as this excellent thesis [Low21]).

3.2.4 Project Ideas: Single-copy, Local Tomography

In the preceding section, we discussed the “textbook" tomography algorithm that is often used in current experimental characterization efforts. As mentioned above, focusing only on the overall parity of the Pauli measurement outcome is inefficient. We obtain $O(n)$ bits of information in each trial of the measurement; however, we only use 1 bit to form our estimator. So, a natural question is whether there exists a Pauli tomography algorithm for which fewer samples suffice.

Optimized Pauli Tomography. In recent years several papers have discussed this issue. An excellent final project would be digging into optimized Pauli tomography in Refs. [Yu20; Ach+25b; Ach+25a].

Experimental Single-setting Tomography. Pauli tomography, as well randomized measurement techniques, require $O(\exp(n))$ measurement *settings*. A very cool paper, Ref. [Str+22] implements a QST protocol using only *one* experimental setting. This is made possible by local symmetric, informationally complete POVMs (SIC-

POVMs). Re-deriving the sample complexity upper bounds from that paper, after an introduction to SIC-POVMs, would make for a great final project.

3.3 Representation Theory Crash Course

> *Have you tried Schur's Lemma?.*
>
> **— Graeme Smith, Comedian and Applied Mathematician**

3.3.1 Warm-up: Haar Averages

Let us denote the unitary group $\mathcal{U}_d \coloneqq \{U \in \mathcal{L}(\mathbb{C}^d) : U^\dagger U = \mathbb{I}_d\}$. The unitary group admits a unique uniform measure called the *Haar measure* which will allow us to take uniform averages over $\mathcal{U}_d$. For a great review of the Haar measure with quantum information theorists in mind, see Ref. [Mel24].

Definition 3.3.1 (Haar Measure on $\mathcal{U}_d$)**.** The Haar measure on the unitary group $\mathcal{U}_d$ is the unique probability measure dU that is left- and right-invariant over $\mathcal{U}_d$. That is, for all integrable functions f and all $V \in \mathcal{U}_d$, we have

$$\int_{\mathcal{U}_d} f(U)dU = \int_{\mathcal{U}_d} f(VU)dU = \int_{\mathcal{U}_d} f(UV)dU. \tag{3.52}$$

We note that for any (measurable) $S \subseteq \mathcal{U}_d$, it follows that $\int_S 1dU \geq 0$ and $\int_{\mathcal{U}_d} 1dU = 1$, which makes the Haar measure a probability measure.

Example 3.3.2 (Haar Measure on $\mathcal{U}_1$)**.** The group $\mathcal{U}_1$ consists of all complex numbers of unit modulus, parameterized as $e^{i\theta}$ with $\theta \in [0, 2\pi)$. The Haar measure is the normalized arc length measure on the unit circle,

$$d\mu(e^{i\theta}) = \frac{d\theta}{2\pi}.$$

Left and right invariance follow immediately: multiplication by a fixed $e^{i\phi}$ shifts $\theta \mapsto \theta + \phi$, which preserves $d\theta$. Integration against this measure is simply averaging over the circle,

$$\int_{\mathcal{U}_1} f \, d\mu = \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta}) \, d\theta.$$

Example 3.3.3 (Haar Measure on Single-Qubit States)**.** The Haar measure on the space of single-qubit pure states $|\psi\rangle \in \mathbb{C}^2$ is the uniform measure on the Bloch sphere,

$$d\mu = \frac{\sin\theta}{4\pi}\, d\theta\, d\phi,$$

where $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$ are the polar and azimuthal angles. This is correctly normalized since

$$\int d\mu = \frac{1}{4\pi}\int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta = \frac{2\pi \cdot 2}{4\pi} = 1.$$

A concrete manifestation of unitary invariance is that averaging the density matrix over all Haar-random states yields the maximally mixed state,

$$\int |\psi\rangle\langle\psi|\, d\mu = \frac{1}{2} I,$$

where $|\psi\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\phi}\sin\frac{\theta}{2}|1\rangle$. Generalizing this to higher dimensions require more advanced tools.

This last example motivates us to define what we mean when we say a state is Haar random.

Definition 3.3.4 (Haar Random State)**.** A *Haar random state* is a state of the form $|\psi\rangle = U|\psi_0\rangle$, where $|\psi_0\rangle \in \mathbb{C}^d$ is a fixed pure state and U is drawn from the Haar measure on $\mathcal{U}_d$. Further, we denote the average over Haar random states as

$$\int_{\psi} f(\psi) d\mu(\psi) = \int_{\mathcal{U}_d} f(U|\psi_0\rangle\langle\psi_0|U^\dagger) d\mu(U) \tag{3.53}$$

In this course, and in quantum information theory generally, we will be interested in computing averages of states, operators, functionals, etc with respect to the Haar measure. To motivate why this necessitates learning a bit of representation theory, let us start by considering a very simple Haar average.

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger] := \int_{U(d)} UOU^\dagger d\mu(U). \tag{3.54}$$

If you already know representation theory, I still encourage you to try and compute this integral using only linear algebra. First, let's try and gain some intuition for what the operator should be when $d = 2$.

Let $A := \mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger]$ denote the Haar average operator we seek. First, let's note that we would need the trace of these to be the same:

$$\begin{aligned}
\operatorname{tr}[A] &= \operatorname{tr}\left[\mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger]\right], && (3.55)\\
&= \operatorname{tr}\left[\int_{U(d)} UOU^\dagger d\mu(U)\right], && (3.56)\\
&= \int_{U(d)} \operatorname{tr}\left[UOU^\dagger\right] d\mu(U), && (3.57)\\
&= \int_{U(d)} \operatorname{tr}[O]\, d\mu(U), \quad \text{cyclicity} && (3.58)\\
&= \operatorname{tr}[O], && (3.59)
\end{aligned}$$

where the last line follows from the normalization of the Haar measure. With this in mind, let us first consider $O = |0\rangle\langle 0|$. In this case, it becomes the projector onto the Haar average pure state. In other words,

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[U|0\rangle\langle 0|U^\dagger] = \int_{\mathcal{U}_2} U|0\rangle\langle 0|U^\dagger d\mu(U) = \int_\psi |\psi\rangle\langle\psi| d\mu(\psi) = \frac{\mathbb{I}}{2}, \tag{3.60}$$

which we may write suggestively as

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[U|0\rangle\langle 0|U^\dagger] = \frac{\operatorname{tr}[|0\rangle\langle 0|]}{2}\cdot\mathbb{I}. \tag{3.61}$$

This is natural: we are averaging all states uniformly, so the resulting state shouldn't point in any preferred direction. Suppose instead we take $O = Z$. Physically, $\langle\psi| Z |\psi\rangle$ corresponds to the length of the projection of $|\psi\rangle$ onto the z-axis of the Bloch sphere. Generally, then, $\langle\psi| UZU^\dagger |\psi\rangle$ is the length of the projection onto the axis that results from rotating Z by some amount. Thus, we would expect the Haar average of such an expectation to be zero.

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[UZU^\dagger] = \mathop{\mathbb{E}}_{U\sim\mu_H}[U|0\rangle\langle 0|U^\dagger] - \mathop{\mathbb{E}}_{U\sim\mu_H}[U|1\rangle\langle 1|U^\dagger] = 0, \tag{3.62}$$

which we can write in a similar form as above $\frac{\operatorname{tr}[Z]}{2}\cdot\mathbb{I}$. Finally, what if $O = \mathbb{I}$? Naturally, we have

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[U\mathbb{I}U^\dagger] = \mathop{\mathbb{E}}_{U\sim\mu_H}[UU^\dagger] = \mathop{\mathbb{E}}_{U\sim\mu_H}[\mathbb{I}] = \mathbb{I}, \tag{3.63}$$

which can be expressed as $\frac{\operatorname{tr}[\mathbb{I}]}{2}\cdot\mathbb{I}$. We conjecture that this form holds in general.

Proposition 3.3.5 (Haar Average of an Operator)**.**

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger] = \frac{\operatorname{tr}[O]}{d}\mathbb{I}. \tag{3.64}$$

Proof. Our conjecture is that this operator is proportional to the identity. Note that the identity is the only unitary matrix that commutes with all other unitary matrices. Thus, if we can show A commutes with all unitaries $V \in U(d)$, we will be able to conclude that A is indeed proportional to the identity. Observe

$$
\begin{aligned}
VAV^\dagger &= V\left(\mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger]\right)V^\dagger, && (3.65)\\
&= V\left(\int_{\mathcal{U}_d} UOU^\dagger d\mu(U)\right)V^\dagger, && (3.66)\\
&= \left(\int_{\mathcal{U}_d} VUOU^\dagger V^\dagger d\mu(U)\right), && (3.67)\\
&= \left(\int_{\mathcal{U}_d} VUO(VU)^\dagger d\mu(U)\right), && (3.68)\\
&= \left(\int_{\mathcal{U}_d} UOU d\mu(U)\right), \quad \text{left-invariance of Haar measure} && (3.69)\\
&= A, && (3.70)
\end{aligned}
$$

which implies $VA = AV$ for all $V \in \mathcal{U}_d$. If the proposition is true, we need to show that this operator is proportional to the identity. A useful fact from linear algebra is that an operator is proportional to the identity *if and only if* it commutes with all other operators. So, if we can show that commuting with all unitaries implies commuting with all operators, we are done. It suffices to show that an operator can be expressed as a linear combination of unitary operators.

To this end [3], note that we can always express an operator in terms of two Hermitian operators

$$
A = \underbrace{\frac{A + A^\dagger}{2}}_{:=H_1} + i \cdot \underbrace{\frac{A - A^\dagger}{2i}}_{:=H_2}. \tag{3.71}
$$

Hermitian operators have real eigenvalues. Further, we can restrict our attention to operators satisfying $\|H\|_\infty$. If this is not the case, simply define $H' = H/\|H\|_\infty$. For such Hermitian operators, $I - H^2 \geq 0$ (i.e. it is a positive operator) and thus it has a unique positive square root $\sqrt{I - H^2}$. From this, we may construct a unitary as

$$
U_\pm := H \pm i\sqrt{I - H^2}, \tag{3.72}
$$

which is manifestly unitary. It follows that $H = (U_+ + U_-)/2$. Doing this for H_1 and H_2 above, we see that we can express arbitrary operators in terms of at most four unitaries. By linearity, it follows that

[3] This construction was inspired by Problem 10 in Sec. 7D of *Linear Algebra Done Right* [Axl24]. I encourage you to think of even more elementary constructions.

$$[A, V] = 0 \; \forall \; V \in \mathcal{U}_d \implies [A, V] = 0 \; \forall \; V \in \mathbb{C}^{d\times d} \implies A = \lambda \mathbb{I}, \tag{3.73}$$

for some $\lambda \in \mathbb{C}$. We can solve for this constant by taking the trace of both sides of this expression

$$\operatorname{tr}\left[\mathop{\mathbb{E}}_{U\sim\mu_H}[UOU^\dagger]\right] = \operatorname{tr}\left[\lambda \mathbb{I}\right], \tag{3.74}$$

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[\operatorname{tr}\left[UOU^\dagger\right]] = \lambda \cdot d, \tag{3.75}$$

$$\mathop{\mathbb{E}}_{U\sim\mu_H}[\operatorname{tr}\left[O\right]] = \lambda \cdot d, \quad \text{cyclicity of trace} \tag{3.76}$$

$$\implies \lambda = \frac{\operatorname{tr}\left[O\right]}{d}, \tag{3.77}$$

as desired. □

This is great! Nothing but some “basic" linear algebra was needed. In many applications; however, we will want to compute higher moments of the so-called moment operator.

Definition 3.3.6 (k-th Moment Operator)**.** The k-th moment operator, with respect to the probability measure μ_H, is defined as $\mathcal{M}^{(k)}_{\mu_H} : \mathcal{L}\left((\mathbb{C}^d)^{\otimes k}\right) \to \mathcal{L}\left((\mathbb{C}^d)^{\otimes k}\right)$:

$$\mathcal{M}^{(k)}_{\mu_H}(O) := \mathop{\mathbb{E}}_{U\sim\mu_H}\left[U^{\otimes k} O U^{\dagger\otimes k}\right], \tag{3.78}$$

for all operators $O \in \mathcal{L}\left((\mathbb{C}^d)^{\otimes k}\right)$.

Quick Quiz 3.3.7. What made the $k = 1$ case tractable using only linear algebra? What breaks down for all $k > 1$.

In the $k = 1$ case, showing that $\mathcal{M}^{(k)}_{\mu_H}(O)$ commutes with all unitaries allowed us to prove that it commutes with all operators (and thus must be proportional to the identity). To introduce some necessary jargon, we say that the *commutant* of $\mathcal{M}^{(1)}_{\mu_H}(O)$ is one dimensional, i.e. spanned by $\mathbb{I}$.

Definition 3.3.8 (Commutant). Given $S \subseteq \mathcal{L}\left(\mathbb{C}^d\right)$, we define its k-th order *commutant* as

$$\text{Comm}(S,k) := \left\{A \in \mathcal{L}\left((\mathbb{C}^d)^{\otimes k}\right) : \left[A, B^{\otimes k}\right] = 0 \; \forall B \in S\right\}. \tag{3.79}$$

As soon as we go to $k = 2$, the commutant is no longer trivial, and it is quite tedious to derive a closed form expression for $\mathcal{M}^{(2)}_{\mu_H}(O)$. Hopefully the strategy is now clear, though. Because the moment operator commutes with the action of the unitary group on $(\mathbb{C}^d)^{\otimes k}$, it lies in the k-th order commutant. Thus, we hope that by classify the k-th order commutant in a sufficiently simple manner, we will be able to express the moment operator in simple terms for all k. To do this, we need some representation theory!

3.3.2 The Church of the Symmetric Subspace

Denote the symmetric group on n elements as $\mathcal{S}_n$. Then, for all $\pi \in \mathcal{S}_n$, define

$$P_d(\pi) \left|i_1\right\rangle \otimes \cdots \otimes \left|i_n\right\rangle = \left|i_{\pi^{-1}(1)}\right\rangle \otimes \cdots \otimes \left|i_{\pi^{-1}(n)}\right\rangle. \tag{3.80}$$

It follows that

$$\begin{aligned}
P_d(\pi)P_d(\sigma) \left|i_1, \ldots, i_n\right\rangle &= P_d(\pi) \left|i_{\sigma^{-1}(1)}, \ldots, i_{\sigma^{-1}(n)}\right\rangle && (3.81)\\
&= P_d(\pi) \left|j_1, \ldots, j_n\right\rangle && (3.82)\\
&= \left|j_{\pi^{-1}(1)}, \ldots, j_{\pi^{-1}(n)}\right\rangle && (3.83)\\
&= \left|i_{\sigma^{-1}\pi^{-1}(1)}, \ldots, i_{\sigma^{-1}\pi^{-1}(n)}\right\rangle, && (3.84)\\
&= \left|i_{(\pi\sigma)^{-1}(1)}, \ldots, i_{(\pi\sigma)^{-1}(n)}\right\rangle, && (3.85)\\
&= P_d(\pi\sigma) \left|i_1, \ldots, i_n\right\rangle, && (3.86)
\end{aligned}$$

where the last line follows from the fact that $(\pi\sigma)^{-1} = \sigma^{-1}\pi^{-1}$, because $(\sigma^{-1}\pi^{-1})(\pi\sigma) = e$. This, along with the fact that $P_d(e) = \mathbb{I}_d$, implies that P_d is a representation of $\mathcal{S}_n$ on $(C^d)^{\otimes n}$. Moreover, it follows that the explicit matrix representation for any $\pi \in \mathcal{S}_n$ can be expressed as

$$P_d(\pi) = \sum_{i_1,\ldots,i_n \in [d]} \left|i_{\pi^{-1}(1)}, \ldots, i_{\pi^{-1}(n)}\right\rangle\!\left\langle i_1, \ldots, i_n\right|. \tag{3.87}$$

Quick Quiz 3.3.9. Show that this representation is unitary.

Taking the adjoint of $P_d(\pi)$ simply swaps bras and kets:

$$P_d(\pi)^\dagger = \sum_{i_1,\ldots,i_n \in [d]} |i_1,\ldots,i_n\rangle \langle i_{\pi^{-1}(1)},\ldots,i_{\pi^{-1}(n)}| . \tag{3.88}$$

Relabeling the dummy summation variables by $j_k = i_{\pi^{-1}(k)}$, i.e. $i_k = j_{\pi(k)}$, this becomes

$$P_d(\pi)^\dagger = \sum_{j_1,\ldots,j_n \in [d]} |j_{\pi(1)},\ldots,j_{\pi(n)}\rangle \langle j_1,\ldots,j_n| = P_d(\pi^{-1}), \tag{3.89}$$

where in the last step we used $(\pi^{-1})^{-1} = \pi$. Unitarity then follows immediately from the homomorphism property,

$$P_d(\pi)^\dagger P_d(\pi) = P_d(\pi^{-1})P_d(\pi) = P_d(\pi^{-1}\pi) = P_d(e) = \mathbb{I}. \tag{3.90}$$

Quick Quiz 3.3.10. Compute the explicit form of $P_d(\pi)$ for all elements of $\mathcal{S}_2$.

The simplest non-trivial case is $\mathcal{S}_2$. Here we can leave the elements alone $\pi = (1)(2)$ or we can swap (transpose) them $\pi = (12)$. The unitary representations of these permutations are the well-known identity and SWAP operators given as

$$\mathbb{I}_d = \sum_{i \in [d]} |i,i\rangle\langle i,i|, \tag{3.91}$$

$$\mathbb{F} = \sum_{i,j \in [d]} |j,i\rangle\langle i,j|. \tag{3.92}$$

For many applications in quantum learning theory, and quantum information theory generally, an object of central importance is the *symmetric subspace*.

Definition 3.3.11 (Symmetric subspace)**.** The *symmetric subspace* of $(\mathbb{C}^d)^{\otimes n}$, denoted $\vee^n \mathbb{C}^d$, is defined to be

$$\vee^n \mathbb{C}^d = \{|\psi\rangle \in (\mathbb{C}^d)^{\otimes n} : P_d(\pi)|\psi\rangle = |\psi\rangle , \quad \forall\, \pi \in \mathcal{S}_n\} \tag{3.93}$$

We can then define the *symmetrizer* as

$$\Pi^{d,n}_{\text{sym}} = \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} P_d(\pi), \tag{3.94}$$

whose name is justified by the following proposition.

Proposition 3.3.12 (Orthogonal Projector onto $\vee^n \mathbb{C}^d$)**.** $\Pi_{\text{sym}}^{d,n}$ is the orthogonal projector onto $\vee^n \mathbb{C}^d$.

Proof. Recall from linear algebra that an operator Π is an orthogonal projector if and only if it satisfies two conditions:

1. *Idempotency*: $\Pi^2 = \Pi$, which ensures that any vector already in the image of Π is left invariant by Π, since if $|v\rangle = \Pi |u\rangle$ then $\Pi |v\rangle = \Pi^2 |u\rangle = \Pi |u\rangle = |v\rangle$.

2. *Hermiticity*: $\Pi^\dagger = \Pi$, which ensures that the image and kernel of Π are orthogonal complements. To see this, note that for any $|u\rangle$ in the image and $|v\rangle$ in the kernel, $\langle u| \Pi^\dagger |v\rangle = \langle u| \Pi |v\rangle = 0$.

These two conditions can be packaged into the single equation

$$\Pi^\dagger \Pi = \Pi, \tag{3.95}$$

since taking the adjoint of both sides gives $\Pi^\dagger \Pi = \Pi^\dagger$, and hence $\Pi = \Pi^\dagger$ (Hermiticity). Substituting this back then yields $\Pi^2 = \Pi$ (idempotency). Thus, to show that the symmetrizer we defined is an orthogonal projector, we may write

$$\begin{aligned}
(\Pi_{\text{sym}}^{d,n})^\dagger \Pi_{\text{sym}}^{d,n} &= \left(\frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} P_d(\pi) \right)^\dagger \Pi_{\text{sym}}^{d,n}, && (3.96)\\
&= \left(\frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} P_d(\pi^{-1}) \right) \Pi_{\text{sym}}^{d,n}, && (3.97)\\
&= \left(\frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} P_d(\pi^{-1}) \right) \left(\frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} P_d(\sigma) \right), && (3.98)\\
&= \frac{1}{(n!)^2} \sum_{\pi \in \mathcal{S}_n} \sum_{\sigma \in \mathcal{S}_n} P_d(\pi^{-1}) P_d(\sigma), && (3.99)\\
&= \frac{1}{(n!)^2} \sum_{\pi \in \mathcal{S}_n} \sum_{\sigma \in \mathcal{S}_n} P_d(\pi^{-1}\sigma), && (3.100)\\
&= \frac{1}{(n!)^2} \sum_{\pi' \in \mathcal{S}_n} \sum_{\sigma \in \mathcal{S}_n} P_d(\pi'), && (3.101)\\
&= \frac{1}{n!} \sum_{\pi' \in \mathcal{S}_n} P_d(\pi'), && (3.102)\\
&= \Pi_{\text{sym}}^{d,n}. && (3.103)
\end{aligned}$$

It remains to show that $\operatorname{Im} \Pi_{\text{sym}}^{d,n} = \vee^n \mathbb{C}^d$. To show $\operatorname{Im} \Pi_{\text{sym}}^{d,n} \subseteq \vee^n \mathbb{C}^d$, observe that for any $|\psi\rangle \in (\mathbb{C}^d)^{\otimes n}$ we have

$$P_d(\pi)\Pi_{\text{sym}}^{d,n} |\psi\rangle = \Pi_{\text{sym}}^{d,n} |\psi\rangle \,, \tag{3.104}$$

thus $\operatorname{Im} \Pi_{\text{sym}}^{d,n} \subseteq \vee^n \mathbb{C}^d$. For the other direction, we observe that if $|\psi\rangle \in \vee^n \mathbb{C}^d$, then

$$\Pi_{\text{sym}}^{d,n} |\psi\rangle = \frac{1}{n!} \sum_{\pi \in \mathcal{S}_n} P_d(\pi) |\psi\rangle = |\psi\rangle \,, \tag{3.105}$$

which establishes $\vee^n \mathbb{C}^d \subseteq \operatorname{Im} \Pi_{\text{sym}}^{d,n}$ and we are done. □

> **Quick Quiz 3.3.13** (Dimension of the symmetric subspace.). What is the dimension of $\vee^n \mathbb{C}^d$?

There are, of course, multiple ways to do this. Perhaps the most familiar technique would be constructing a basis for the space and then counting the number of elements. The main observation is that a basis for the symmetric subspace can be obtained by taking a standard basis for $(\mathbb{C}^d)^{\otimes n}$ and then symmetrizing. Because symmetrizing is just averaging over all permutations, the resulting basis should not depend on the order of the elements, but only the number of times a certain element occurs in the decomposition. To determine the number of distinct elements in this basis, then we ask how many ways are there to place n indistinguishable items in d distinguishable bins. Next, observe that d bins can be viewed as $d-1$ bars between n identical stars (borrowing the standard language from an introductory discrete math course). For example, $d = 3$ and $n = 4$ is shown below:

$$\star\star\,|\,\star\,|\,\star \tag{3.106}$$

How many arrangements of these symbols are there? Well, there are $n + d - 1$ symbols and we are to select n to be stars (the rest are, consequently, bars). The total number of such choices is the number of distinct basis elements of the symmetric subspace, yielding

$$\dim \vee^n \mathbb{C}^d = \binom{n+d-1}{n} = \frac{(n+d-1)!}{n!(d-1)!}. \tag{3.107}$$

See Exercise for an alternative technique.

Having defined the permutation operators, we may state a result from representation theory that is of central importance to quantum learning theory.

Theorem 3.3.14 (Schur-Weyl Duality). The k-th order commutant of the unitary group is the span of the permutation operators associated to S_k:

$$\text{Comm}(\mathcal{U}_d, k) = \text{span}\{P_d(\pi) : \pi \in \mathcal{S}_k\}. \tag{3.108}$$

This is a standard result in representation theory which we will not prove here (see Refs. [Mel24; Har13; Led25] for more details). It is worth noting that one direction of the proof is simple: $\text{span}\{P_d(\pi) : \pi \in \mathcal{S}_k\} \subseteq \text{Comm}(\mathcal{U}_d, k)$. To see why this is true, consider an arbitrary permutation $P_d(\pi)$ with $\pi \in \mathcal{S}_k$ and $U \in \mathcal{U}_d$. We have:

$$\begin{aligned}
P_d(\pi)U^{\otimes k} |\psi_1\rangle \otimes \cdots \otimes |\psi_k\rangle &= P_d(\pi)(U|\psi_1\rangle) \otimes \cdots \otimes (U|\psi_k\rangle) && (3.109)\\
&= P_d(\pi) |\phi_1\rangle \otimes \cdots \otimes |\phi_k\rangle, \quad |\phi_j\rangle := U|\psi_j\rangle && (3.110)\\
&= |\phi_{\pi^{-1}(1)}\rangle \otimes \cdots \otimes |\phi_{\pi^{-1}(k)}\rangle && (3.111)\\
&= U|\psi_{\pi^{-1}(1)}\rangle \otimes \cdots \otimes U|\psi_{\pi^{-1}(k)}\rangle && (3.112)\\
&= U^{\otimes k} P_d(\pi) |\psi_1\rangle \otimes \cdots \otimes |\psi_k\rangle, && (3.113)
\end{aligned}$$

for all $|\psi_1\rangle, \ldots, |\psi_k\rangle \in \mathbb{C}^d$. Hence we have that $\left[P_d(\pi), U^{\otimes k}\right] = 0$ for all $\pi \in \mathcal{S}_k$, so indeed $\text{span}\{P_d(\pi) : \pi \in \mathcal{S}_k\} \subseteq \text{Comm}(\mathcal{U}_d, k)$.

Theorem 3.3.15 (Computing Moments, Thm. 10 [Mel24]). Let $O \in \mathcal{L}\left((\mathbb{C}^d)^{\otimes k}\right)$. The moment operator can then be expressed as a linear combination of permutation operators:

$$\mathop{\mathbb{E}}_{U\sim\mu_H}\left[U^{\otimes k} O U^{\dagger\otimes k}\right] = \sum_{\pi\in\mathcal{S}_k} c_\pi(O) P_d(\pi), \tag{3.114}$$

where the coefficients $c_\pi(O)$ can be determined by solving the following linear system of $k!$ equations:

$$\text{Tr}\left(P_d^\dagger(\sigma) O\right) = \sum_{\pi\in\mathcal{S}_k} c_\pi(O)\, \text{Tr}\left(P_d^\dagger(\sigma) P_d(\pi)\right) \quad \text{for all } \sigma \in \mathcal{S}_k. \tag{3.115}$$

This system always has at least one solution.

See Ref. [Mel24] for a proof of this result. To get a feel for this result, let us apply it when $k = 2$.

Quick Quiz 3.3.16. Find a closed form for

$$\mathop{\mathbb{E}}_{U\sim\mu_H}\left[U^{\otimes 2} O U^{\dagger\otimes 2}\right]. \tag{3.116}$$

For $k = 2$, the symmetric group $\mathcal{S}_2$ has two elements: the identity e and the swap (12). So by the theorem, the moment operator takes the form:

$$\mathop{\mathbb{E}}_{U\sim\mu_H}\left[U^{\otimes 2}OU^{\dagger\otimes 2}\right] = c_e(O)P_d(e) + c_{(12)}(O)P_d((12)) = c_e(O)\mathbb{I} + c_{(12)}(O)\mathbb{F}, \tag{3.117}$$

where $\mathbb{F}$ denotes the swap operator. The linear system gives two equations, one for $\sigma = e$ and one for $\sigma = (12)$:

$$\mathrm{tr}\left[O\right] = c_e(O)\mathrm{tr}\left[\mathbb{I}\right] + c_{(12)}(O)\mathrm{tr}\left[\mathbb{F}\right], \tag{3.118}$$

$$\mathrm{tr}\left[\mathbb{F}^\dagger O\right] = c_e(O)\mathrm{tr}\left[\mathbb{F}\right] + c_{(12)}(O)\mathrm{tr}\left[\mathbb{F}^\dagger\mathbb{F}\right]. \tag{3.119}$$

Using the identities $\mathrm{tr}\left[\mathbb{I}\right] = d^2$, $\mathrm{tr}\left[\mathbb{F}\right] = d$, and $\mathrm{tr}\left[\mathbb{F}^\dagger\mathbb{F}\right] = d^2$, the system becomes:

$$\mathrm{tr}\left[O\right] = d^2 c_e(O) + d\, c_{(12)}(O), \tag{3.120}$$

$$\mathrm{tr}\left[\mathbb{F}O\right] = d\, c_e(O) + d^2 c_{(12)}(O). \tag{3.121}$$

Solving this 2×2 linear system gives:

$$c_e(O) = \frac{d\mathrm{tr}\left[O\right] - \mathrm{tr}\left[\mathbb{F}O\right]}{d(d^2-1)}, \tag{3.122}$$

$$c_{(12)}(O) = \frac{d\mathrm{tr}\left[\mathbb{F}O\right] - \mathrm{tr}\left[O\right]}{d(d^2-1)}. \tag{3.123}$$

Substituting back, the closed form for the second moment operator is:

$$\mathop{\mathbb{E}}_{U\sim\mu_H}\left[U^{\otimes 2}OU^{\dagger\otimes 2}\right] = \frac{d\mathrm{tr}\left[O\right] - \mathrm{tr}\left[\mathbb{F}O\right]}{d(d^2-1)}\mathbb{I} + \frac{d\mathrm{tr}\left[\mathbb{F}O\right] - \mathrm{tr}\left[O\right]}{d(d^2-1)}\mathbb{F}. \tag{3.124}$$

Quick Quiz 3.3.17. Evaluate the following integral

$$\int_\psi \psi^{\otimes 2} d\psi. \tag{3.125}$$

The goal is for students to realize that this integral over Haar random states can be mapped to the form we just saw above, with $O = |\psi\rangle\langle\psi|^{\otimes 2}$. Using Def. 3.3.4, we may write

$$\int_\psi \psi^{\otimes 2} d\psi = \int_{\mathcal{U}_d} U^{\otimes 2}|0\rangle\langle 0|(U^\dagger)^{\otimes 2} d\mu(U), \tag{3.126}$$

$$= \mathop{\mathbb{E}}_{U\sim\mu_H}[U^{\otimes 2}|0\rangle\langle 0|^{\otimes 2}(U^\dagger)^{\otimes 2}], \tag{3.127}$$

$$= \frac{1}{d(d+1)}\left(\mathbb{I} + \mathbb{F}\right), \tag{3.128}$$

where we have used Eq. (3.124) and the fact that the trace of a quantum state is one.

3.3.3 Continuous POVMs

Thus far in this course, we have only considered measurements corresponding to a discrete set of POVM elements. In the next few sections, we will utilize continuous POVMs in our tomography algorithms. So, we pause briefly to introduce the relevant notions. This treatment is adopted almost directly from Michael Walter's excellent lecture notes on symmetry and quantum information [Wal18]. I highly recommend exploring those notes in detail, but here we will only need a few fundamental definitions[4].

Definition 3.3.18 (Continuous POVM)**.** Let Ω be a measurable outcome space. Then, a *continuous POVM* on a Hilbert space $\mathcal{H}$ is defined by a collection of operators $\{Q_x\}_{x\in\Omega}$ and an associated measure dx satisfying

1. **Positivity.** $Q_x \geq 0$ for all $x \in \Omega$,

2. **Completeness.** $\int_\Omega Q_x dx = \mathbb{I}_{\mathcal{H}}$.

The *probability density* of the outcome distribution with respect to the measure dx (Born's rule) is given by

$$p_\psi(x) = \mathrm{tr}\left[Q_x |\psi\rangle\langle\psi|\right]. \tag{3.129}$$

From this definition, it follows that probabilities and expectation values can be computed as

$$\Pr_\psi[\text{outcome} \in S] = \int_S \mathrm{tr}\left[Q_x |\psi\rangle\langle\psi|\right] dx, \tag{3.130}$$

$$\mathbb{E}_\psi[f(x)] = \int \mathrm{tr}\left[Q_x |\psi\rangle\langle\psi|\right] f(x) dx, \tag{3.131}$$

where $S \subseteq \Omega$ is some measurable subset of all possible outcomes. The discrete POVMs we saw in Def. 2.1.2 can be recovered as a special case of this more general result. If Ω is finite, then one may take the counting measure $dx(S) := |S|$ and identify $\int_S dx = \sum_{x\in S}$.

[4] If you are interested in a measure-theoretic treatment of quantum mechanics, I highly recommend Frederic Schuller's excellent courses on YouTube or in scribed lecture notes [Sch15].

Quick Quiz 3.3.19. When we discussed Naimark's theorem, we said we needed an ancillary space large enough to store all possible measurement outcomes. So, can we every hope to implement a continuous POVM in finite dimensions?

Amazingly, the answer is that yes, one can do this. In a classic paper, Chiribella *et al.* [CDS07]

> "*establish a fundamental property of quantum measurements with a continuous set of outcomes, namely, that for finite-level systems any such measurement is equivalent to a continuous random choice of measurements with a* ***finite*** *number of outcomes. This means that any physical quantity measured on a finite dimensional system is intrinsically discrete, while the continuum is pure classical randomness.*"

I highly encourage the interested reader to take a look at this paper. For now, we will accept that a continuous POVM can be implemented (although not necessarily in a time-efficient manner) and will proceed with our study of tomography algorithms.

3.3.4 Exercises

Exercise 3.3.1 (Alternative Proof of Symmetric Subspace Dimension)**.** Use Prop 3.3.12 to prove that

$$\dim \vee^n \mathbb{C}^d = \binom{n+d-1}{n}. \tag{3.132}$$

3.4 Single-copy, Global Tomography Algorithms

Suppose, for example, that the unknown state we are trying to learn was diagonal in some basis. We would just measure in that basis and output the result as our guess. In expectation, this would yield the correct state. However, because we do not know this preferred basis, we might as well guess randomly.

3.4.1 Tomography via Uniform POVM

Consider a POVM defined in terms of the measurement operators $Q_\psi = d \cdot |\psi\rangle\langle\psi|$, where $|\psi\rangle$ is sampled with respect to the Haar measure $d\psi$. Clearly, $Q_\psi \geq 0$ and

$$\int_\psi d|\psi\rangle\langle\psi|d\psi = d \cdot \frac{1}{d}\mathbb{I} = \mathbb{I}, \tag{3.133}$$

so the set $\{Q_\psi\}$ forms a valid continuous POVM. This measurement is referred to as the *uniform POVM* in the literature. Let's see how we can use it to construct a single-copy, global tomography algorithm.

Algorithm 3 Uniform POVM Tomography

Require: d-dimensional state ρ, number of copies n.

1: **for** $i = 1$ **to** n **do**
2: Measure the uniform POVM $\{Q_\psi\}$ on a single copy of ρ.
3: Receive outcome $|\psi_i\rangle$.
4: $\hat{\rho}_i \leftarrow (d+1)|\psi_i\rangle\langle\psi_i| - \mathbb{I}_d$
5: **end for**
6: **return** $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}_i$

The following proposition motivates the form of the estimator constructed in this algorithm.

Proposition 3.4.1 (Expected output of uniform POVM)**.** Suppose we measure a state ρ with the uniform POVM. Then, the expected value of the output is given as

$$\mathbb{E}\left[|\psi\rangle\langle\psi|\right] = \int_\psi \mathrm{tr}\left[Q_\psi \rho\right] |\psi\rangle\langle\psi| d\psi = \frac{1}{d+1}\left(\mathbb{I} + \rho\right). \tag{3.134}$$

Proof. We may write

$$\mathbb{E}\left[|\psi\rangle\langle\psi|\right] = \int_{\psi} \mathrm{tr}\left[Q_{\psi}\rho\right]|\psi\rangle\langle\psi| d\psi, \tag{3.135}$$

$$= d \cdot \int_{\psi} \mathrm{tr}\left[|\psi\rangle\langle\psi|\rho\right]|\psi\rangle\langle\psi| d\psi, \tag{3.136}$$

$$= d \cdot \int_{\psi} \mathrm{tr}_B[(\mathbb{I}\otimes\rho)\cdot|\psi\rangle\langle\psi|\otimes|\psi\rangle\langle\psi|] d\psi, \quad \mathrm{tr}\left[M\right]\cdot A = \mathrm{tr}_B[A\otimes M] \tag{3.137}$$

$$= d \cdot \mathrm{tr}_B[(\mathbb{I}\otimes\rho)\cdot\left(\int_{\psi}|\psi\rangle\langle\psi|^{\otimes 2} d\psi\right)], \quad \text{linearity} \tag{3.138}$$

$$= d \cdot \mathrm{tr}_B[(\mathbb{I}\otimes\rho)\cdot\frac{1}{d(d+1)}(\mathbb{I}\otimes\mathbb{I}+\mathbb{F})], \quad \text{Eq. (3.128)} \tag{3.139}$$

$$= \frac{1}{d+1}\left(\mathbb{I}+\rho\right), \tag{3.140}$$

where, in the last line we have used the identity $\mathrm{tr}_B[\mathbb{I}\otimes\rho] = I\mathrm{tr}\left[\rho\right] = \mathbb{I}$ as well as the fact that $\mathrm{tr}[\mathbb{F}(\mathbb{I}\otimes\rho)] = \rho$. This result, which may be dubbed the "partial SWAP trick," is worth working through, so we have left it as Exercise 3.4.1.

□

3.4.2 Project Idea: Tomography Via Unitary 2-designs

3.4.3 Project Idea: Optimal Pure State Tomography via Unitary 2-designs

3.4.4 Exercises

Exercise 3.4.1 (Partial SWAP trick)**.** Let ρ be a quantum state on $\mathbb{C}^d$ and $\mathbb{F}$ the swap operator on $\mathbb{C}^d\otimes\mathbb{C}^d$. Then, show

$$\mathrm{tr}_B[\mathbb{F}(\mathbb{I}\otimes\rho)] = \rho. \tag{3.141}$$

3.5 Multi-copy, Global Tomography Algorithms

3.5.1 Pure State Tomography via Uniform POVM

3.5.2 Mixed State Tomography Reduces to Pure State Tomography

3.6 Lower Bounds on Sample Complexity

Having now seen several algorithms for quantum state tomography, we turn to the natural question: are they optimal? In other words, can we prove that no QST algorithm could ever use fewer samples than the algorithms we have constructed. On the surface, it sounds much more complicated to prove such a no-go result. Although the techniques differ greatly across quantum learning theory, the high-level logic is ubiquitous.

To prove a lower bound on the sample complexity of a learning or testing task, a common strategy is to use a *reduction* from state discrimination to the learning/testing task of interest. Reducing problem A to problem B formally just means that if we have an algorithm for problem B, we could use it to solve problem A. Thus, a lower bound on the complexity of problem A implies a lower bound on the complexity of problem B. In our case, we will essentially be reducing state discrimination to tomography with sufficient accuracy. Let's jump in!

We will follow the treatment in Angus Lowe's Master's thesis [Low21]. **Add literature review/history.**

3.6.1 Universal Lower Bound

The last algorithm we studied for QST reduced mixed state tomography to pure state tomography and used $O(d^2/\epsilon^2)$ samples in the worst case (i.e. when ρ was full rank). The natural question we will answer in this section is whether or not $\Omega(d^2/\epsilon^2)$ samples are *necessary* for this task in the worst case.

> **Theorem 3.6.1** (Universal Lower Bound on QST)**.** Given access to $\rho^{\otimes n}$, any algorithm outputting $\hat{\rho}$ such that $d_{\mathrm{tr}}(\rho, \hat{\rho}) \leq \epsilon$ with at least constant probability of success must use at least $n = \Omega(d^2/\epsilon^2)$ samples of the state.

The proof idea will be to show that there exists a large set of states that are all separated enough to allow a tomographic procedure (with sufficient accuracy) to distinguish them with high probability. This is the reduction of state discrimination to tomography. We will then show that, unless $n = \Omega(d^2/\epsilon^2)$, we would violate fundamental information theoretic inequalities.

As a conceptual tool, we can map this distinguishing task to a communication protocol. This will make the connection to the information theoretic quantities feel much more natural. Suppose Alice and Bob have agreed to encode some 2^M quantum states into bitstrings of length M. Then, Alice might send $\rho_x^{\otimes n}$ to Bob for some message $x \in \{0,1\}^M$. If Bob can perform QST with sufficient accuracy, he could decode this message. Through this, albeit impractical communication protocol, Alice would have successfully communicated M bits of information.

Now, let N denote the cardinality of the set of well-separated states corresponding to distinct messages. Further, let x be a uniform random variable over $[N]$ and y denote the outcome of a measurement on the state $\rho_x^{\otimes n}$. Then, the mutual information $I(x:y)$ between these two random variables will be upper and lower bounded via Holevo's and Fano's inequality, respectively

$$\Omega(\log N) \leq I(x:y) \leq n\epsilon^2. \tag{3.142}$$

Quick Quiz 3.6.2. What do these inequalities imply about the requirements on the set of possible state? How must N scale in order to yield our desired $n = \Omega(d^2/\epsilon^2)$ lower bound?

Clearly, these inequalities cannot both be true unless $n = \Omega(\log N/\epsilon^2)$. Thus, we need $N = \exp\left(\Omega(d^2)\right)$ to yield our desired bound! Can we really construct a set with doubly-exponentially many quantum states that are sufficiently well-separated to be distinguished? At first glance, this may seem infeasible but fortunately, as Carlton Caves once said "Hilbert space is a big place."

Let us start by showing that we can construct such a large set of quantum states. Consider states of the form

$$\rho_{\epsilon,U} := \epsilon U\sigma U^\dagger + (1-\epsilon)\frac{\mathbb{I}}{d}, \tag{3.143}$$

where $\sigma := \frac{2}{d}Q$, for a fixed rank-$\frac{d}{2}$ orthogonal projector and $\epsilon \in (0,1)$.

Quick Quiz 3.6.3. Show that this forms a valid quantum state.

Verifying this requires recalling two facts from linear algebra: 1) orthogonal projectors are Hermitian and 2) the trace of an orthogonal projector is equal to its rank. It follows immediately, then, that $\rho_{\epsilon,U} \geq 0$ (i.e. it is PSD) and that $\mathrm{tr}\left[\rho_{\epsilon,U}\right] = 1$. Thus, it is a valid state. With this form of state in mind, we may prove the first crucial lemma.

Quick Quiz 3.6.4. Determine the spectrum of $\rho_{\epsilon,U}$.

We can write $\rho_{\epsilon,U} = \frac{2\epsilon}{d} UQU^\dagger + (1-\epsilon)\frac{\mathbb{I}}{d}$. Now, because Q is an orthogonal projector, we know the eigenvalues can only be zero or one. Thus, Q has $d/2$ non-zero eigenvalues all equal to 1. Further, a change of basis does not change the spectrum, thus, $UQU^\dagger$ has the same spectrum. With these facts in mind, we can show that $\rho_{\epsilon,U}$ has $d/2$ eigenvalues equal to $(1+\epsilon)/d$ and $d/2$ eigenvalues equal to $(1-\epsilon)/d$.

Proposition 3.6.5. There exists a universal constant c such that the following holds. Pick $\epsilon \in (0,1)$, let $d > 0$ be a positive integer, and let $0 \leq N < \frac{1}{2}e^{cd^2}$ be an integer. Consider the set of states $\{\rho_1, \rho_2, \ldots, \rho_N\} \subset D(d)$ where

$$\rho_i := \epsilon U_i \sigma U_i^\dagger + (1-\epsilon)\frac{\mathbb{I}}{d}, \tag{3.144}$$

for each $i \in [N]$, $U_1, U_2, \ldots, U_N \in U(d)$ are arbitrary unitary operators and σ is as in Eq. (3.143). For Haar-random $\boldsymbol{U}$ taking values in $U(d)$, the probability that $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$ for some $i \in [N]$ is strictly less than 1. That is,

$$\operatorname*{Prob}_{\boldsymbol{U}\sim\mu_H}\left[\bigcup_{i=1}^{N}\left\{\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \frac{\epsilon}{2}\right\}\right] < 1. \tag{3.145}$$

To motivate this lemma, let us see how it will be used.[5]

Definition 3.6.6 (ϵ-packing condition)**.** A set of mixed states $\mathcal{S}$ satisfies the ϵ-packing condition for some $\epsilon > 0$ if it holds that $\|\rho - \rho'\|_1 > \epsilon$ for every $\rho, \rho' \in \mathcal{S}$ such that $\rho \neq \rho'$.

Corollary 3.6.7. Let $d > 0$ be a positive integer. There exists a set of $N \geq \exp(\Omega(d^2))$ quantum states as in Lemma 3.6.5 for which the $(\epsilon/2)-$packing condition is satisfied.

Proof. Our goal is to find the largest lower bound on N. Let us start from the most trivial lower bound and use induction to steadily improve the bound. Suppose we start with $k = 1$ states of the form given in Lemma 3.6.5, that is a set $\mathcal{S}_1 = \{\rho_1\}$ which is trivially an $\epsilon/2$-packing. By the lemma, we know that the probability of a Haar random state $\rho_{\epsilon,\boldsymbol{U}}$ being within $\epsilon/2$ of this state is strictly less than one. Thus there must exist some state we can add to our packing. The existence of this next state is guaranteed as long as $k < e^{cd^2 - \ln(2)}$ with the same c from the Lemma. Now

[5] Note that this proof, and many of the details in this section, were originally written by Luke Coffman while he was an undergraduate working with the author.

suppose $\mathcal{S}_k = \{\rho_1, \ldots, \rho_k\} \subset D(d)$ with $k < e^{cd^2 - \ln(2)}$ satisfies the packing condition, that is $\|\rho_i - \rho_j\|_1 > \epsilon/2$ for all $i, j \in [k]$, $i \neq j$. Then from the lemma we have that the probability of choosing a unitary $\boldsymbol{U}$ Haar randomly such that $\mathcal{S}_k \cup \{\rho_{\epsilon,\boldsymbol{U}}\}$ no longer satisfies the $\epsilon/2$-packing condition is strictly less than one. Therefore, there must exist at least one state which we can add to the packing. By performing induction on k we guarantee the existence of an $\epsilon/2$ packing for $1 \leq k \leq e^{cd^2 - \ln(2)}$, where the right inequality is no longer strict since we can add one state just below and saturate. Note that beyond this upper bound on k we can no longer make any existence claims, this does not imply no states can be added to the packing, it just implies no states can be added with this method. This leads us to choose the largest lower bound of $N \geq \exp(\Omega(d^2))$. To see this scaling explicitly, we have shown the existence of a packing of size $N = \frac{1}{2}e^{cd^2}$. Taking logarithms gives $\ln N = cd^2 - \ln 2$, and since $\ln 2$ is a constant, we have $\ln N = \Omega(d^2)$, or equivalently $N = \exp(\Omega(d^2))$, where the implicit constant in the Ω depends only on the universal constant c. □

Now, in order to prove Proposition 3.6.5, we need the following lemmas regarding the concentration of measure for functions of Haar random matrices. But first, a quick definition!

Definition 3.6.8 (Lipschitz Continuity)**.** Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is said to be *L-Lipschitz* if there exists a constant $L \geq 0$ such that for all $x, x' \in X$,

$$d_Y(f(x), f(x')) \leq L \cdot d_X(x, x'). \tag{3.146}$$

In the specific case relevant here, we take $X = U(d)$ equipped with the metric induced by the Frobenius norm (i.e. the Schatten 2-norm), $Y = \mathbb{R}$ equipped with the standard metric, so that the condition becomes: for all $U, V \in U(d)$,

$$|f(U) - f(V)| \leq L\|U - V\|_2. \tag{3.147}$$

Lemma 3.6.9 (Concentration of measure)**.** Let $f : \mathbb{U}(d) \to \mathbb{R}$ be an L-Lipschitz function with respect to the metric induced by the Frobenius norm, and let $\mu := \mathbb{E}_{\boldsymbol{U} \sim \mu_H} f(\boldsymbol{U})$. Then, for any $t > 0$, it holds that

$$\Pr_{\boldsymbol{U} \sim \mu_H}[|f(\boldsymbol{U}) - \mu| \geq t] \leq 2\exp\left(-\frac{dt^2}{12L^2}\right) \tag{3.148}$$

Lemma 3.6.10 (Concentration of projector overlaps)**.** Let $\boldsymbol{U}$ be a Haar-random unitary operator taking values in $\mathcal{U}_d$ and let $P, Q \in PSD(d)$ be orthogonal projectors with rank r_P, r_Q respectively. It holds that

$$\Pr_{\boldsymbol{U}\sim\mu_H}\left[\left|\operatorname{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right] - \frac{r_Q r_P}{d}\right| \geq t\right] \leq 2\exp\left(-\frac{cdt^2}{\sqrt{r_P r_Q}}\right), \tag{3.149}$$

where c is some universal constant.

Proof. Let $f : \mathbb{U}(d) \to \mathbb{R}$ be defined as $f(U) = \operatorname{tr}\left[PUQU^\dagger\right]$ for all $U \in \mathbb{U}(d)$. We will show that the expectation of the function is $\frac{r_P r_Q}{d}$ and that the Lipschitz constant is $\mathcal{O}((r_P r_Q)^{1/4})$. For the expectation, we have

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right]\right] &= \int_{\mathcal{U}_d} \operatorname{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right] d\mu(U), && (3.150)\\
&= \operatorname{tr}\left[\int_{\mathcal{U}_d} P\boldsymbol{U}Q\boldsymbol{U}^\dagger d\mu(U)\right], \quad \text{linearity of trace} && (3.151)\\
&= \operatorname{tr}\left[P\int_{\mathcal{U}_d} \boldsymbol{U}Q\boldsymbol{U}^\dagger d\mu(U)\right], && (3.152)\\
&= \operatorname{tr}\left[P\frac{\operatorname{tr}[Q]}{d}\mathbb{I}\right], \quad \text{Lemma 3.3.5} && (3.153)\\
&= \frac{\operatorname{tr}[Q]\operatorname{tr}[P]}{d}, && (3.154)\\
\mathbb{E}\left[\operatorname{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right]\right] &= \frac{r_Q r_P}{d}, && (3.155)
\end{aligned}$$

as desired. Now, recall that a function is L-Lipschitz (with respect to the metric induced by the Frobenius norm) if there exists $L \geq 0$ such that for all $U, V \in U(d)$

$$|f(U) - f(V)| \leq L\|U - V\|_2. \tag{3.156}$$

To find L, we may write

$$\begin{aligned}
|f(U) - f(V)| &= \left|\operatorname{tr}\left[PUQU^\dagger\right] - \operatorname{tr}\left[PVQV^\dagger\right]\right|, && (3.157)\\
&= \left|\operatorname{tr}\left[P(UQU^\dagger - VQV^\dagger)\right]\right|, && (3.158)\\
&= \left|\frac{1}{2}(\operatorname{tr}\left[P(UQU^\dagger - VQV^\dagger)\right] + \operatorname{tr}\left[P(UQU^\dagger - VQV^\dagger)\right])\right|, && (3.159)\\
&= \frac{1}{2}\left|\operatorname{tr}\left[P(U+V)Q(U-V)^\dagger\right] + \operatorname{tr}\left[P(U-V)Q(U+V)^\dagger\right]\right|, && (3.160)\\
&\leq \frac{1}{2}\left|\operatorname{tr}\left[P(U+V)Q(U-V)^\dagger\right]\right| + \frac{1}{2}\left|\operatorname{tr}\left[P(U-V)Q(U+V)^\dagger\right]\right|, && (3.161)
\end{aligned}$$

which follows from the triangle inequality. We can then upper bound each term in this sum. It will be useful to note that Cauchy-Schwarz takes the following form for the Hilbert-Schmidt inner product

$$\left|\mathrm{tr}\left[A^\dagger B\right]\right| \leq \sqrt{\mathrm{tr}\left[A^\dagger A\right]}\sqrt{\mathrm{tr}\left[B^\dagger B\right]} = \|A\|_2\|B\|_2, \tag{3.162}$$

and that the Schatten 2-norm is unitarily invariant, i.e.

$$\begin{aligned}
\|UA\|_2 &= \sqrt{\mathrm{tr}\left[UA(UA)^\dagger\right]}, && (3.163)\\
&= \sqrt{\mathrm{tr}\left[UAA^\dagger U^\dagger\right]}, && (3.164)\\
&= \sqrt{\mathrm{tr}\left[AA^\dagger\right]}, \quad \text{cyclicity of trace} && (3.165)\\
&= \|A\|_2. && (3.166)
\end{aligned}$$

With these facts in mind, we may upper bound the first term in Eq. (3.161) as follows

$$\begin{aligned}
&\frac{1}{2}\left|\mathrm{tr}\left[P(U+V)Q(U-V)^\dagger\right]\right| && (3.167)\\
&= \frac{1}{2}\left|\mathrm{tr}\left[PUQ(U-V)^\dagger + PVQ(U-V)^\dagger\right]]\right|, && (3.168)\\
&\leq \frac{1}{2}\left|\mathrm{tr}\left[PUQ(U-V)^\dagger\right]\right| + \frac{1}{2}\left|\mathrm{tr}\left[PVQ(U-V)^\dagger\right]\right|, \quad \text{triangle inequality,} && (3.169)\\
&\leq \frac{1}{2}\left(\|PUQ\|_2\|U-V\|_2\right) + \frac{1}{2}\left(\|PVQ\|_2\|U-V\|_2\right), \quad \text{CS inequality,} && (3.170)\\
&= \frac{1}{2}\left(\|PUQ\|_2 + \|PVQ\|_2\right)\|U-V\|_2, && (3.171)\\
&= \frac{1}{2}\left(\sqrt{\mathrm{tr}\left[PUQ(PUQ)^\dagger\right]} + \sqrt{\mathrm{tr}\left[PVQ(PVQ)^\dagger\right]}\right)\|U-V\|_2, && (3.172)\\
&= \frac{1}{2}\left(\sqrt{\mathrm{tr}\left[PUQU^\dagger P)\right]} + \sqrt{\mathrm{tr}\left[PVQV^\dagger P\right]}\right)\|U-V\|_2, \quad Q, P \text{ Hermiticity} && (3.173)\\
&= \frac{1}{2}\left(\sqrt{\mathrm{tr}\left[PUQU^\dagger)\right]} + \sqrt{\mathrm{tr}\left[PVQV^\dagger\right]}\right)\|U-V\|_2, \quad \text{cyclicity of trace} && (3.174)\\
&\leq \frac{1}{2}\left(\sqrt{\|P\|_2\|Q\|_2} + \sqrt{\|P\|_2\|Q\|_2}\right)\|U-V\|_2, \quad \text{CS + unitary invariance} && (3.175)\\
&= \sqrt{\|P\|_2\|Q\|_2}\|U-V\|_2, && (3.176)\\
&= (r_P r_Q)^{1/4}\|U-V\|_2. && (3.177)
\end{aligned}$$

Essentially the same steps can be followed to arrive at an overall bound of

$$|f(U) - f(V)| \leq 2(r_P r_Q)^{1/4}\|U-V\|_2, \tag{3.178}$$

which implies the desired Lipschitz constant of $L = 2(r_P r_Q)^{1/4}$. □

Proof of Proposition 3.6.5. Let $P := \mathbb{I} - Q$ be the orthogonal projector associated with the fixed rank-$\frac{d}{2}$ projector given in Eq. (3.143). Note that because Q is Hermitian

and rank-$\frac{d}{2}$, so is P. Finally, note that $P - Q$ is unitary as one can check from the definition. With these facts in mind, we can take $t = \frac{d}{8}$ in Lemma 3.6.10 to see

$$\Pr_{\boldsymbol{U}\sim\mathcal{U}_d}\left[\left|\mathrm{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right] - \frac{d}{4}\right| \geq \frac{d}{8}\right] \leq 2\exp\left(-cd^2\right). \tag{3.179}$$

Then, because $\Pr[|x| \geq a] = \Pr[x \leq -a] + \Pr[x \geq a]$, we have $\Pr[|x| \geq a] \geq \Pr[x \leq -a]$, we have

$$\Pr_{\boldsymbol{U}\sim\mathcal{U}_d}\left[\mathrm{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right] - \frac{d}{4} \leq -\frac{d}{8}\right] \leq \Pr_{\boldsymbol{U}\sim\mathcal{U}_d}\left[\left|\mathrm{tr}\left[P\boldsymbol{U}Q\boldsymbol{U}^\dagger\right] - \frac{d}{4}\right| \geq \frac{d}{8}\right] \leq 2\exp\left(-cd^2\right). \tag{3.180}$$

Then, consider the quantity

$$\begin{aligned}
\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{I}} &= \epsilon U\sigma U^\dagger + (1-\epsilon)\frac{\mathbb{I}}{d} - (\epsilon\sigma + (1-\epsilon)\frac{\mathbb{I}}{d}), && (3.181)\\
&= \epsilon(U\sigma U^\dagger - \sigma), && (3.182)\\
&= \frac{2\epsilon}{d}(UQU^\dagger - Q), && (3.183)
\end{aligned}$$

where the first equality follows from Eq. 3.143. Next, note that for any $A \in H(d)$ and $V \in U(d)$, we have from the spectral decomposition of A

$$\begin{aligned}
|\mathrm{tr}\left[AV\right]| &= \left|\sum_i a_i \left\langle a_i\right| V \left|a_i\right\rangle\right|, && (3.184)\\
&\leq \sum_i |a_i| \left|\left\langle a_i\right| V \left|a_i\right\rangle\right|, && \text{triangle} \quad (3.185)\\
&\leq \sum_i |a_i|, && \text{normalized vectors} \quad (3.186)\\
&= \|A\|_1. && (3.187)
\end{aligned}$$

Applying this fact by right-multiplying by $P - Q$, we find

$$\|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{I}}\|_1 = \left\|\frac{2\epsilon}{d}(UQU^\dagger - Q)\right\|_1, \tag{3.188}$$

$$\geq \frac{2\epsilon}{d}\left|\mathrm{tr}\left[(UQU^\dagger - Q)(P - Q)\right]\right|, \tag{3.189}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P - \underbrace{QP}_{=0} - UQU^\dagger Q + Q\right]\right|, \tag{3.190}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + \mathrm{tr}\left[-UQU^\dagger Q + Q\right]\right|, \tag{3.191}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + \mathrm{tr}\left[Q - QUQU^\dagger\right]\right|, \tag{3.192}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + \mathrm{tr}\left[Q(\mathbb{I} - UQU^\dagger)\right]\right|, \tag{3.193}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + \mathrm{tr}\left[(\mathbb{I} - P)(\mathbb{I} - UQU^\dagger)\right]\right|, \tag{3.194}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + \mathrm{tr}\left[\mathbb{I}\right] - \mathrm{tr}\left[P\right] - \mathrm{tr}\left[UQU^\dagger\right] + \mathrm{tr}\left[PUQU^\dagger\right]\right|, \tag{3.195}$$

$$= \frac{2\epsilon}{d}\left|\mathrm{tr}\left[UQU^\dagger P\right] + d - \frac{d}{2} - \frac{d}{2} + \mathrm{tr}\left[PUQU^\dagger\right]\right|, \tag{3.196}$$

$$= \frac{4\epsilon}{d}\left|\mathrm{tr}\left[PUQU^\dagger\right]\right|. \tag{3.197}$$

Now, if we force the distance between these states to be $\epsilon/2$, we have

$$\frac{\epsilon}{2} \geq \|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{I}}\|_1 \geq \frac{4\epsilon}{d}\left|\mathrm{tr}\left[PUQU^\dagger\right]\right|, \tag{3.198}$$

which in turn implies for any Haar random unitary, $\boldsymbol{U}$ we have

$$\Pr_{\boldsymbol{U}\sim\mathcal{U}_d}\left[\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{I}}\|_1 \leq \frac{\epsilon}{2}\right] \leq 2\exp\left(-cd^2\right). \tag{3.199}$$

Finally, consider some $i \in [N]$ and a specific U_i and ρ_i as defined above. We have that for any Haar-random unitary $\boldsymbol{U}$ that

$$\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{I}}\|_1 = \|\rho_{\epsilon,U_i\boldsymbol{U}} - \rho_i\|_1, \qquad \text{unitary invariance of trace distance} \tag{3.200}$$

$$= \|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1, \qquad \text{left-invariance of Haar,} \tag{3.201}$$

which implies that

$$\Pr_{\boldsymbol{U}\sim\mathcal{U}_d}\left[\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \frac{\epsilon}{2}\right] \leq 2\exp\left(-cd^2\right). \tag{3.202}$$

Then, because this inequality holds for all $i \in [N]$, we can use the union bound to obtain

$$\Pr_{U\sim\mathcal{U}_d}\left[\cup\{\|\rho_{\epsilon,U}-\rho_i\|_1 \leq \frac{\epsilon}{2}\}\right] \leq \sum_{i=1}^{N} \Pr_{U\sim\mathcal{U}_d}\left[\|\rho_{\epsilon,U}-\rho_i\|_1 \leq \frac{\epsilon}{2}\right], \tag{3.203}$$

$$\leq \sum_{i=1}^{N} 2\exp\left(-cd^2\right), \tag{3.204}$$

$$= 2N\exp\left(-cd^2\right). \tag{3.205}$$

To ensure this probability is strictly less than 1, then, we must take

$$2N\exp\left(-cd^2\right) < 1 \implies N < \frac{1}{2}\exp\left(cd^2\right). \tag{3.206}$$

□

Let's take stock of what we have accomplished. Using a probabilistic existence argument, we have shown that there must exist an $\epsilon/2$-packing with $N = \exp\left(\Omega(d^2)\right)$ states. In addition to being excellent support for the claim that "Hilbert Space is a big place," it will allow us to prove a tight universal lower bound on the sample complexity of QST.

To proceed, we will need a few results from quantum information theory. If these are completely unfamiliar, please pause here and read about them in Refs. [NC00; Wil17] before continuing.

Theorem 3.6.11 (Holevo's Theorem)**.** Let $\{\rho_x\}_{x=1}^N$ be quantum states prepared with probabilities p_x, and let $\{M_y\}$ be any POVM. Define X to be the classical random variable with distribution p_x and Y to be the outcome of the measurement. Then, the mutual information between X and Y is upper bounded as

$$I(X:Y) \leq S\left(\sum_x p_x\rho_x\right) - \sum_x p_x S(\rho_x), \tag{3.207}$$

where $S(\rho) = -\mathrm{tr}(\rho\log\rho)$ is the von Neumann entropy.

Quick Quiz 3.6.12. What states minimize and maximize the von Neumann entropy?

Even if you have never seen this particular formula, perhaps you've heard of entropy in the context of thermodynamics as a measure of "disorder" or in classical statistics

as a measure of "surprisal." As such, we might guess that pure states should have zero entropy and maximally mixed states should have maximal entropy. This intuition is correct, as the following calculation shows.

Pure states. Let $\rho = |\psi\rangle\langle\psi|$ be a pure state. Then ρ has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = \cdots = \lambda_d = 0$, so

$$S(\rho) = -\sum_{i=1}^{d} \lambda_i \log \lambda_i = -1 \cdot \log 1 - \sum_{i=2}^{d} 0 \cdot \log 0 = 0, \tag{3.208}$$

where we use the convention $0 \log 0 = 0$.

Maximally mixed state. Let $\rho = \frac{\mathbb{I}}{d}$. Then ρ has eigenvalues $\lambda_i = \frac{1}{d}$ for all $i \in [d]$, so

$$S\left(\frac{\mathbb{I}}{d}\right) = -\sum_{i=1}^{d} \frac{1}{d} \log \frac{1}{d} = -d \cdot \frac{1}{d} \log \frac{1}{d} = -\log \frac{1}{d} = \log d. \tag{3.209}$$

Let's apply this theorem to our problem at hand.

Lemma 3.6.13. Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\}$ be an $\epsilon/2$-packing with cardinality $N = \exp \Omega(d^2)$ as in Corollary 3.6.7. Let X be uniformly random over $[N]$ and consider the ensemble of states corresponding to $\rho_x^{\otimes n}$. The Holevo information for this ensemble satisfies

$$S\left(\frac{1}{N}\sum_{i=1}^{N} \rho_i^{\otimes n}\right) - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i^{\otimes n}) \leq n\epsilon^2. \tag{3.210}$$

Proof. Let's start by letting $\rho := \frac{1}{N}\sum_{i=1}^{N} \rho_i^{\otimes n}$ and noting that the reduced state on any of the n individual registers is $\tau := \frac{1}{N}\sum_{i=1}^{N} \rho_i$. The proof of this Lemma requires several facts about entropies. First, the Von Neumann entropy is subadditive. That is, for an n-partite system with state $\rho_{A_1 A_2 \cdots A_n}$, the subadditivity of von Neumann entropy states:

$$S(A_1 A_2 \cdots A_n) \leq \sum_{i=1}^{n} S(A_i), \tag{3.211}$$

where $S(A_i) = S(\rho_{A_i})$ is the von Neumann entropy of the reduced state. Equality holds if and only if $\rho_{A_1 A_2 \cdots A_n} = \rho_{A_1} \otimes \rho_{A_2} \otimes \cdots \otimes \rho_{A_n}$, i.e. the global state is a product state. In our context, this implies that $S(\rho) \leq nS(\tau)$ and $S(\rho_i^{\otimes n}) = nS(\rho_i)$. Thus, it suffices to show that

$$S(\tau) - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i) \leq \epsilon^2. \tag{3.212}$$

Well, the first term is just the entropy of a mixed quantum state which is always upper bounded by $\log d$. What about the second? For this, we could just discard the sum because it is non-negative, but we want a tighter, ϵ-dependent bound. Recall that ρ_i's are of the form

$$\rho_i := \frac{2\epsilon}{d} U_i Q U_i^\dagger + (1-\epsilon)\frac{\mathbb{I}}{d}, \tag{3.213}$$

thus half of the eigenvalues are equal to $(1-\epsilon)/d$ and the other half are equal to $(1+\epsilon)/d$. It follows that the von Neumann entropy is

$$\begin{aligned} S(\rho_i) &= -\frac{d}{2}\frac{1+\epsilon}{d}\log\left(\frac{1+\epsilon}{d}\right) - \frac{d}{2}\frac{1-\epsilon}{d}\log\left(\frac{1-\epsilon}{d}\right), && (3.214)\\ &= -\frac{1+\epsilon}{2}\left(\log\left(\frac{1+\epsilon}{2}\right) + \log\left(\frac{2}{d}\right)\right) - \frac{1-\epsilon}{2}\left(\log\left(\frac{1-\epsilon}{2}\right) + \log\left(\frac{2}{d}\right)\right), && (3.215)\\ &= H\left(\frac{1+\epsilon}{2}\right) + \log\left(\frac{d}{2}\right), && (3.216) \end{aligned}$$

where we have introduced the so-called *binary entropy function* $H(p) = -p\log p - (1-p)\log(1-p)$. With this in mind we may return to the expression we were attempting to bound

$$\begin{aligned} S(\tau) - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i) &\leq \log d - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i), && (3.217)\\ &= \log d - \log d/2 - H\left(\frac{1+\epsilon}{2}\right), && (3.218)\\ &\leq 1 - (1-\epsilon^2), && (3.219)\\ \implies S(\tau) - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i) &\leq \epsilon^2, && (3.220) \end{aligned}$$

as desired. In the penultimate line, we used a common lower bound on the binary entropy (see Exercise 3.6.1). This concludes the proof. □

Okay, we now have that $I(X:Y) \leq n\epsilon^2$. To conclude the proof, we need to find a strong lower bound on the mutual information. To achieve this goal, we will use another workhorse of classical information theory.

Lemma 3.6.14 (Fano's Inequality)**.** Let X, Y, Z be discrete random variables forming a Markov chain $X \to Y \to Z$, where X takes values in $\mathcal{X}$. It holds that

$$H(p_e) + p_e \log(|\mathcal{X}|) \geq H(X|Y), \tag{3.221}$$

where $p_e := \Pr[X \neq Z]$ and $H(\cdot)$ is the binary entropy function.

Quick Quiz 3.6.15. Let X, Y, Z be discrete random variables forming a Markov chain $X \to Y \to Z$. Suppose Alice has a message X which is uniformly random over N distinct values, and Bob is able to decode the message with constant probability of success using Z. Show that

$$I(X:Y) = \Omega(\log(N)). \tag{3.222}$$

Proof. Using the definition of mutual information we have $I(X:Y) = H(X) - H(X|Y)$. Let p_e be as in Lemma 3.6.14. By Lemma 3.6.14 we have $I(X:Y) \geq H(X) - p_e \log(N) - H(p_e)$. Using the fact that $H(X) = \log(N)$ for uniformly random X and $H(p_e) \leq 1$ we obtain

$$I(X:Y) \geq (1 - p_e)\log(N) - 1. \tag{3.223}$$

Since Bob decodes with constant probability of success, $p_e \leq 1 - c$ for some constant $c \in (0, 1]$, and therefore $1 - p_e \geq c > 0$. Thus:

$$I(X:Y) \geq c\log(N) - 1 = \Omega(\log(N)), \tag{3.224}$$

where the last step holds since the -1 is a constant and $c > 0$ is a constant, so the expression grows as $\log(N)$ as $N \to \infty$. □

Putting all of the pieces together, we have that

$$\Omega(\log N) \leq I(X:Y) \leq n\epsilon^2, \tag{3.225}$$

which can only be true if we take $n = \Omega(\log N/\epsilon^2)$. But $N = \exp(\Omega(d^2))$, so, $\log n = \Omega(d^2)$, and altogether we have

$$n = \Omega\left(\frac{d^2}{\epsilon^2}\right). \tag{3.226}$$

This is a beautiful result. Moreover, it is tight, as we saw there exists a mixed state tomography algorithm that, in the worst case, uses $n = O(d^2/\epsilon^2$ samples. Thus, the sample complexity of mixed state tomography is

$$n = \Theta\left(\frac{d^2}{\epsilon^2}\right). \tag{3.227}$$

There is much more that could be said about quantum state tomography in theory and in practice; however, we will end here and mention many exciting possible projects realted to QST in the next section.

3.6.2 Project Ideas: QST Lower Bounds

I will add more detail about each of these, but want to list the papers for now:

- Optimal lower bounds for quantum state tomography [SSW25].
- Lower Bounds for Learning Quantum States with Single-Copy Measurements [LN25]

3.6.3 Exercise

Exercise 3.6.1. Let $H(p) = -p\log p - (1-p)\log(1-p)$ be the binary entropy function. Show that for $|p - 1/2| \leq \delta$,

$$H(p) \geq 1 - 4\delta^2. \tag{3.228}$$

4 Quantum Shadow Tomography

> *Measure now, ask questions later!*
>
> **— Steve Flammia?**

4.1 Warm-up: Direct Observable Estimation

4.2 Classical Shadow Tomography

I may add my own derivations of the classical shadow formalism, but for now, please see Robert or Jordan/Sitan's course notes for a really nice treatment of the framework.

The punchline is that we can use the same estimator for the state that we derived in Sec. 3

$$\hat{\rho}_i = (d+1)U^\dagger |b_i\rangle\langle b_i| U_i - \mathbb{I}. \tag{4.1}$$

Because this is an unbiased estimator of the state, i.e. $\mathbb{E}\left[\hat{\rho}_i\right] = \rho$, it follows from linearity that

$$X_i = \mathrm{tr}\left[O\hat{\rho}_i\right] \tag{4.2}$$

is an unbiased estimator of the expectation values we wish to estimate. That is,

$$\mathbb{E}\left[X_i\right] = \mathbb{E}\left[\mathrm{tr}\left[O\hat{\rho}_i\right]\right] = \mathrm{tr}\left[O\mathbb{E}\left[\hat{\rho}_i\right]\right] = \mathrm{tr}\left[\rho O\right]. \tag{4.3}$$

I will add the variance analysis at some point (again, see Robert's notes) but for now, it suffices to recall that we could show the following state-independent bound on the variance of these estimators

$$\mathrm{Var}\left[X_i\right] \leq 3\mathrm{tr}\left[O^2\right]. \tag{4.4}$$

It is claimed that this is a surprising bound due to the fact that it does not depend on the state or n directly; however, it can in general be very loose. Before dwelling on this point, let us finish the story.

The claimed power of the shadow formalism is that one can estimate many observable expectation values even if those observables are determined after the data collection phase. The following theorem is the bedrock of the classical shadow formalism.

Theorem 4.2.1 (Classical Shadow Tomography, [HKP20]). Let $O_1, \ldots, O_M$ be a set of M observables, and let $B := \max_i \operatorname{tr}[O_i^2]$. To predict the expectation value $\operatorname{tr}[O_i\rho]$ for all M observables simultaneously from a single set of N classical shadows, up to an additive error ϵ and with a total success probability of at least $1-\delta$, it suffices to use

$$N = O\left(\frac{B\log(M/\delta)}{\epsilon^2}\right) \tag{4.5}$$

samples and only single-copy measurements on ρ.

Proof. Using the median-of-means estimator for a single observable, we showed that

$$N = O\left(\frac{\operatorname{tr}[O_i^2]}{\epsilon^2}\log\frac{1}{\delta'}\right) \tag{4.6}$$

measurements suffice to estimate a single observable to ϵ precision with probability of failure at most δ'. The worst case is the O_i which maximizes the Frobenius norm. Thus, at worst,

$$N = O\left(\frac{B}{\epsilon^2}\log\frac{1}{\delta'}\right) \tag{4.7}$$

measurements suffice. To ensure that all M predictions are ϵ-accurate, we must use a union bound

$$\Pr[\text{at least one failure}] = \Pr[\cup_{i=1}^{M}\text{failure}_i] \leq \sum_{i=1}^{M}\Pr[\text{failure}_i] \leq M\delta' =: \delta, \tag{4.8}$$

where we have set the overall probability of failure to δ. To achieve this, then, we simply set the worst-case single-observable probability of error to $\delta' = \delta/M$. Plugging this back into the original result, we obtain the desired sample complexity upper of

$$N = \mathcal{O}\left(\frac{B}{\epsilon^2}\log\left(\frac{M}{\delta}\right)\right). \tag{4.9}$$

□

Now, above we showed that if we naively estimated M observables (all satisfying $\|O\|_\infty \leq 1$) by successively estimating each one individually, we would need to use

$$N = O\left(\frac{M}{\epsilon^2}\log\left(\frac{M}{\delta}\right)\right). \tag{4.10}$$

Thus, by using randomized measurements we 1) do not need to be told the observables of interest before we collect data and 2) we can use *exponentially* fewer copies in terms of the number of observables. However, the scaling of classical shadows still could be exponential if the Frobenius norm scales exponentially in the number of qubits.

4.2.1 Interlude: Single-copy Lower Bounds via Le Cam

There are many interesting questions one could ask (and hopefully answer) about the classical shadow formalism. The one to which we now turn is whether we could ever hope to learn all n-qubit Pauli observables without incurring an exponential overhead. That is, we know that Theorem 4.2.1, along with the fact that $\mathrm{tr}\left[P^2\right] = 2^n$ for all n-qubit Paulis, implies that

$$N = O\left(\frac{2^n}{\epsilon^2}\log\left(\frac{M}{\delta}\right)\right) \tag{4.11}$$

copies suffice to learn M Pauli observables. The construction of such an algorithm does not preclude the existence of an algorithm that can achieve this task using fewer copies. In this section, we will introduce a formalism for proving lower bounds against all, potentially adaptive, single-copy lower bounds. This technique was formalized in Ref. [Che+22]; however, the clearest treatment I have seen is in Lecture 21 (Chapter 14) of Ref. [CC25].

Insert more history, literature, and discussion of set-up

The authors refer to algorithms based on potentially adaptive, single-copy measurements as “classical", “conventional," or say that they use no “quantum memory." I think all of these terms are readily misunderstood, so let us very carefully state the assumptions we make about our access model. First, we note that all POVMs on a finite-dimensional quantum system can be simulated by rank-1 POVMs.

Lemma 4.2.2 (Simulating POVMs with Rank-1 POVMs)**.** Consider an arbitrary POVM $\{F_i\}$ with $F_i \geq 0$ and $\sum_i F_i = \mathbb{I}$. The probability distribution $\mathrm{tr}\left[F_i\rho\right]$ can be simulated by rank-1 POVMs on ρ along with classical post-processing.

Proof. Because all POVM elements are positive semi-definite, they may be expressed as

$$F_i = d \sum_j a_{ij} |\phi_{ij}\rangle\langle\phi_{ij}|, \tag{4.12}$$

where $a_{ij} \geq 0$ and where the factor of d ensure that $\{a_{ij}\}$ forms a probability distribution, which will be a convenient choice of normalization. To simulate the distribution, we simply implement the rank-1 POVM and group terms accordingly. That is

$$\sum_j \Pr[(i,j)] = \sum_j \operatorname{tr}\left[d a_{ij} |\phi_{ij}\rangle\langle\phi_{ij}|\rho\right] = \operatorname{tr}\left[\left(\sum_j d a_{ij} |\phi_{ij}\rangle\langle\phi_{ij}|\right)\rho\right] = \operatorname{tr}\left[F_i \rho\right] \tag{4.13}$$

□

Thus, without loss of generality, we will assume that the POVMs we implement are rank 1. With this in mind, we formally define the single-copy access model.

Definition 4.2.3 (Single-copy Access Model)**.** Fix an unknown state ρ on $\mathcal{H} \simeq \mathbb{C}^d$ and let $\{d a_i |\phi_i\rangle\langle\phi_i|\}_i$ denote a rank-1 POVM. The *single-copy access model* is as follows:

1. Prepare fresh copy of ρ and perform rank-1 POVM, $\{d a_i |\phi_i\rangle\langle\phi_i|\}_i$, to obtain outcome $i = q$ which is stored in classical memory.

2. Prepare fresh copy of ρ and perform rank-1 POVM, $\{d a_{q,i} |\phi_{q,i}\rangle\langle\phi_{q,i}|\}_{q,i}$, to obtain outcome $i = r$ which is stored in classical memory.

3. Prepare fresh copy of ρ and perform rank-1 POVM, $\{d a_{q,r,i} |\phi_{q,r,i}\rangle\langle\phi_{q,r,i}|\}_{q,r,i}$, to obtain outcome $i = s$ which is stored in classical memory.

4. Repeat a total of T rounds.

Crucially, the POVM applied in each round may depend on all the outcomes previously observed. That is, the POVMs may be chosen adaptively in each round of the experiment.

A classical learner is defined as an agent that, in each round of a learning algorithm, may implement a POVM (which may depend on all prior measurement outcomes) to obtain classical data that is then stored in classical memory. Because the classical

memory will depend, in general, on all prior measurement outcomes, such learning algorithms lend themselves naturally to a *learning tree representation* which can be defined formally as follows.

Definition 4.2.4 (Learning Tree Representation, Def. 5.1 [Che+22]). Let ρ be an unknown n-qubit quantum state. A classical learning algorithm can be represented as a rooted tree $\mathcal{T}$ of depth T satisfying the following properties:

- Each node v of $\mathcal{T}$ corresponds to a probability $p^{\rho}_{\mathcal{T}}(v)$.
- The root node occurs with unit probability, $p^{\rho}_{\mathcal{T}}(r) = 1$.
- For every non-leaf node u, fix a rank-1 POVM on $\mathcal{H}$ of the form

$$\{da_v |\phi_v\rangle\langle\phi_v|\}_{v\in\text{child}(v)}, \text{ s.t. } a_v \geq 0 \text{ and } \sum_{v\in\text{child}(v)} da_v |\phi_v\rangle\langle\phi_v| = \mathbb{I},$$

  where each $v \in \text{child}(v)$ corresponds to one possible outcome of the POVM performed at node u.
- If v us a child of u, then the probability of transitioning from u to v when the underlying state is ρ is

$$p^{\rho}_{\mathcal{T}}(v) = p^{\rho}_{\mathcal{T}}(u) da_v \text{tr}\left[\rho |\phi_v\rangle\langle\phi_v|\right]. \tag{4.14}$$

- Every root-to-leaf path is of length T. We denote the probability that the classical memory is in state ℓ after T rounds as $p^{\rho}_{\mathcal{T}}(\ell)$.

Add figure to illustrate the tree formalism.

Let $v_0 := r$ and $\ell := v_T$ and note that in a decision tree, a leaf *uniquely determines* the root-to-leaf path through the tree. Thus, by specifying ℓ, one is actually specifying an entire sequence $\boldsymbol{\ell} = (v_0, v_1, \ldots, v_T)$. Moreover, because we prepare an independent copy of ρ in each round of our experiment, the probability that we traverse a specific root-to-leaf path is given simply as

$$p^{\rho}_{\mathcal{T}}(\ell) = \prod_{t=1}^{T} da_{v_t} \text{tr}\left[\rho |\phi_{v_t}\rangle\langle\phi_{v_t}|\right] = d^T \cdot a_{\boldsymbol{\ell}} \cdot \text{tr}\left[\rho^{\otimes T} \bigotimes_{t=1}^{T} |\phi_{v_t}\rangle\langle\phi_{v_t}|\right], \tag{4.15}$$

where we have defined $a_{\boldsymbol{\ell}} := \prod_{t=1}^{T} a_{v_t}$. Now, to distinguish ρ from σ given this access model, one must be able to distinguish $p^{\rho}_{\mathcal{T}}(\ell)$ from $p^{\sigma}_{\mathcal{T}}(\ell)$. Thus, once the measurements are implemented, we are left with a purely classical distinguishing

task. This was covered tangentially in Sec. 2, but we will formalize things a bit more now (See Chapter 14 of Ref. [CC25] for proofs of the following facts).

Recall from Def. 2.2.4 that the total variation distance between two probability distributions p, q is defined as

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2} \cdot \sum_i |p_i - q_i|. \tag{4.16}$$

From this definition, one can show that

$$d_{\mathrm{TV}}(p,q) = \sum_{i \in A} p_i - q_i =: p_A - q_A, \tag{4.17}$$

where $A := \{i : p_i \geq q_i\}$ and where we have defined $p_A := \sum_{i \in A} p_i$ (similarly for q). With this result in mind, we may state the following lemma.

Lemma 4.2.5 (Lemma 225 [CC25]). Suppose $p_i > 0$ for all i, and that the **likelihood ratio** $\frac{q_i}{p_i}$ satisfies $\frac{q_i}{p_i} \geq 1 - c$ for all i and for some constant $c \in [0,1]$. Then $d_{\mathrm{TV}}(p,q) \leq c$.

Proof. Let $A = \{i : p_i \geq q_i\}$. Using the fact that $d_{\mathrm{TV}}(p,q) = p_A - q_A$, together with $\frac{q_i}{p_i} \geq 1 - c$, we have

$$d_{\mathrm{TV}}(p,q) = \sum_{i \in A}(p_i - q_i) = \sum_{i \in A} p_i \left(1 - \frac{q_i}{p_i}\right) \leq c \sum_{i \in A} p_i \leq c \sum_i p_i = c.$$

□

Lastly, letting $S \subseteq \Omega$ be a subset of all possible outcomes, one can show

$$d_{\mathrm{TV}}(p,q) = \sup_S |p_S - q_S|. \tag{4.18}$$

All of these expressions will come in handy, so I encourage you to take the time to prove them! With these facts in hand, we may state the lemma[1] that gives the total variation distance its operational meaning.

Lemma 4.2.6 (Le Cam's Two-point Method, Lemma 223 [CC25]). Let p, q be two probability distributions on a finite set Ω. Suppose we are given a single sample $X \in \Omega$ which is drawn from p with probability $1/2$ and from q with probability $1/2$. For any (possibly randomized) decision rule $\mathcal{A} : \Omega \to \{0,1\}$

[1] We cite Ref. [CC25] because they have a very clear proof of the lemma in the context in which we will utilize it. Note, however, that this technique has been ubiquitous in classical statistical learning theory for several decades.

that outputs a guess for which distribution was used, the success probability satisfies

$$\Pr[\mathcal{A} \text{ correct}] \leq \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(p, q) = \frac{1}{2} + \frac{1}{4} \sum_{i \in \Omega} |p_i - q_i|.$$

Moreover, this bound is tight: there exists a decision rule that achieves equality.

Proof. Let 0 denote the guess that the sample came from p and 1 the guess that the sample came from q. Let $\mathcal{A} : \Omega \to \{0, 1\}$ denote any algorithm used to decide which was the case. Further, let $S \subseteq \Omega$ be the set of outcomes on which $\mathcal{A}$ guesses that the sample came from p, i.e.

$$S := \{i \in \Omega : \mathcal{A}(i) = 0\}.$$

It follows that on S^c the rule guesses that the sample came from q. When the true distribution is p, the rule is correct with probability p_S; when the true distribution is q, it is correct with probability $q_{S^c} = 1 - q_S$. Since each case occurs with prior probability $1/2$, the overall success probability is

$$\begin{aligned}
\Pr[\mathcal{A} \text{ correct}] &= \frac{1}{2} p_S + \frac{1}{2} q_{S^c}, \\
&= \frac{1}{2} p_S + \frac{1}{2} (1 - q_S), \\
&= \frac{1}{2} + \frac{1}{2} (p_S - q_S), \\
&\leq \frac{1}{2} + \frac{1}{2} |p_S - q_S|, \\
&\leq \frac{1}{2} + \frac{1}{2} \sup_S |p_S - q_S|, \\
&= \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(p, q).
\end{aligned}$$

To see that this bound is tight, choose a subset $S^* \subseteq \Omega$ that attains the supremum in the characterization

$$d_{\mathrm{TV}}(p, q) = \sup_S |p_S - q_S|.$$

Define $\mathcal{A}$ so that it guesses p on S^* and q on $(S^*)^c$. For this rule,

$$\Pr[\mathcal{A} \text{ correct}] = \frac{1}{2} + \frac{1}{2} (p_{S^*} - q_{S^*}) = \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(p, q),$$

as claimed. □

Lemma 4.2.7 (Lemma 224, [CC25]). Fix a learning tree $\mathcal{T}$ and a property testing instance with probability distributions $\mu_1, \ldots, \mu_k$ over quantum states on $\mathcal{H}$. For each $j \in [k]$ let

$$p_j(\ell) := \mathbb{E}_{\rho \sim \mu_j} \left[p_{\mathcal{T}}^{\rho}(\ell) \right]$$

denote the induced probability distribution over leaves ℓ of $\mathcal{T}$ when the unknown state is sampled from μ_j. Fix two indices $i, j \in [k]$, and suppose we are promised that the unknown index is either i or j, each with prior probability $1/2$. Then any conventional experiment described by $\mathcal{T}$ that attempts to decide whether the index is i or j has success probability at most

$$\frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}}(p_i, p_j) = \frac{1}{2} + \frac{1}{4} \sum_{\ell} |p_i(\ell) - p_j(\ell)|.$$

In particular, if $d_{\mathrm{TV}}(p_i, p_j) \leq \delta$ for some $\delta \in [0, 1]$, then no such experiment can distinguish between the two hypotheses with success probability greater than $1/2 + \delta/2$.

Assuming there exists a method of successfully distinguishing two distributions, we have $\Omega(1) \leq p_{\mathrm{succ}}$. Coupled with the previous lemma, we have

$$\Omega(1) \leq p_{\mathrm{succ}} \leq \frac{1}{2} + \frac{1}{2} d_{\mathrm{TV}} \left(\mathbb{E}_{\rho \sim \mu_i} \left[p_{\mathcal{T}}^{\rho}(\ell) \right], \mathbb{E}_{\rho \sim \mu_j} \left[p_{\mathcal{T}}^{\rho}(\ell) \right] \right), \tag{4.19}$$

$$\implies \Omega(1) \leq d_{\mathrm{TV}} \left(\mathbb{E}_{\rho \sim \mu_i} \left[p_{\mathcal{T}}^{\rho}(\ell) \right], \mathbb{E}_{\rho \sim \mu_j} \left[p_{\mathcal{T}}^{\rho}(\ell) \right] \right). \tag{4.20}$$

Now, as an illustrative example, suppose we can prove an upper bound on the TV distance of the form $T/f(n)$. Then, we conclude that

$$\Omega(1) \leq \frac{T}{f(n)} \implies T \leq \Omega(f(n)). \tag{4.21}$$

Thus, to prove strong lower bounds, one needs to construct ensembles of quantum states that, via a single-copy measurement protocol, result in distributions over leaf nodes that are statistically close unless the number of samples T is sufficiently large.

4.2.2 Lower Bound on Pauli Shadow Tomography with Single-copy Measurements

This framework will become more clear through an application to Pauli shadow tomography.

Theorem 4.2.8 (Single-copy Pauli Shadow Lower Bound, Cor. 5.9 [Che+22]). Any learning algorithm operating in the single-copy access model requires

$$T \geq \Omega(2^n/\epsilon^2) \tag{4.22}$$

copies of ρ to predict $\{\mathrm{tr}\,[P\rho]\}$ for all Paulis P to at most ϵ-error with high probability.

Proof. First, consider the set of $2(4^n - 1)$ observables $\mathcal{P} := \mathcal{P}_+ \cup \mathcal{P}_-$, where

$$\mathcal{P}_+ := \{\sigma_1 \otimes \cdots \otimes \sigma_n : \sigma_i \in \{\mathbb{I}, X, Y, Z\}\ \forall i\} \setminus \{\mathbb{I}^{\otimes n}\}, \tag{4.23}$$
$$\mathcal{P}_- := \{-\sigma_1 \otimes \cdots \otimes \sigma_n : \sigma_i \in \{\mathbb{I}, X, Y, Z\}\ \forall i\} \setminus \{-\mathbb{I}^{\otimes n}\}, \tag{4.24}$$

and we order the list of observables such that $O_i = -O_{i+M/2}$ for all $i \in [M/2]$. Moreover, note that, for any $O_i \in \mathcal{P}$, we have $\mathrm{tr}\,[O_i^2] = 2^n$ and $\mathrm{tr}\,[O_i] = 0$.

With these definitions in mind, consider the many-versus-one distinguishing task with null and alternative hypotheses given as

$$\rho_0 = \frac{\mathbb{I}}{2^n}, \tag{4.25}$$
$$\rho_i = \frac{\mathbb{I} + 3\epsilon O_i}{2^n}, \quad \forall i \in [M], \tag{4.26}$$

where, in the second case, we imagine sampling a random state uniformly from this ensemble. Next, note that because we have assumed that all of our observables are traceless and have eigenvalues ± 1, we have that $\mathrm{tr}\,[O_i] = 0$ and $\mathrm{tr}\,[O_i^2] = 2^n$, which together imply that $\mathrm{tr}\,[O_i\rho_i] = 3\epsilon$ and $\mathrm{tr}\,[O_i\rho_0] = 0$ for all $i \in [M]$. Thus, for any ρ_i, there exists an observable with $\mathrm{tr}\,[O_i\rho_i]$ much larger than zero, while $\mathrm{tr}\,[O_i\rho_0] = 0$ for all possible $i \in [M]$. Thus, if we can learn these expectation values up to ϵ error with probability at least $2/3$, we can solve the many-versus-one distinguishing task. Thus, a lower bound on the many-versus-one distinguishing task will also be a lower bound on shadow tomography.

To ease notation, let the probability distribution over leaf notes for the null and alternatives be denoted as follows

$$p_{\mathrm{null}}(\ell) := p_{\mathcal{T}}^{\mathbb{I}/2^n}(\ell), \tag{4.27}$$
$$p_{\mathrm{alts}}(\ell) := \mathbb{E}_{\rho\sim\mu}[p_{\mathcal{T}}^{\rho}(\ell)], \tag{4.28}$$

where μ is simply the ensemble of all ρ_i's from which we sample uniformly. Recalling Lemma 4.2.7, we know that it suffices to prove

$$\frac{p_{\text{alts}}(\ell)}{p_{\text{null}}(\ell)} \geq 1 - \delta \tag{4.29}$$

for all $\ell \in \text{leaf}(\mathcal{T})$. Using Eq. (4.15) for the root-to-leaf probability, we may write

$$p_{\text{null}}(\ell) = d^T \cdot a_{\boldsymbol{\ell}} \cdot \text{tr}\left[\left(\frac{\mathbb{I}}{d}\right)^{\otimes T} \bigotimes_{t=1}^{T} |\phi_{v_t}\rangle\langle\phi_{v_t}|\right] = a_{\boldsymbol{\ell}}, \tag{4.30}$$

$$p_{\text{alts}}(\ell) = \mathbb{E}_i\left[d^T \cdot a_{\boldsymbol{\ell}} \cdot \text{tr}\left[\left(\frac{\mathbb{I} + 3\epsilon O_i}{d}\right)^{\otimes T} \bigotimes_{t=1}^{T} |\phi_{v_t}\rangle\langle\phi_{v_t}|\right]\right], \tag{4.31}$$

$$= a_{\boldsymbol{\ell}} \cdot \mathbb{E}_i\left[\text{tr}\left[\left(\mathbb{I} + 3\epsilon O_i\right)^{\otimes T} \bigotimes_{t=1}^{T} |\phi_{v_t}\rangle\langle\phi_{v_t}|\right]\right], \tag{4.32}$$

$$= a_{\boldsymbol{\ell}} \cdot \mathbb{E}_i\left[\prod_{t=1}^{T} 1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right]. \tag{4.33}$$

Our goal, then, is to bound their ratio away from one. We may write

$$\frac{p_{\text{alts}}(\ell)}{p_{\text{null}}(\ell)} = \mathbb{E}_i\left[\prod_{t=1}^{T} 1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right], \tag{4.34}$$

$$= \mathbb{E}_i\left[\exp\log\left(\prod_{t=1}^{T} 1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right)\right], \tag{4.35}$$

$$= \mathbb{E}_i\left[\exp\sum_{t=1}^{T} \log\left(1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right)\right], \tag{4.36}$$

$$\geq \exp\sum_{t=1}^{T} \mathbb{E}_i\left[\log\left(1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right)\right], \quad \text{Jensen's Inequality} \tag{4.37}$$

$$= \exp\sum_{t=1}^{T} \frac{1}{M}\sum_{i=1}^{M} \log\left(1 + 3\epsilon \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle\right), \tag{4.38}$$

$$= \exp\sum_{t=1}^{T} \frac{1}{M}\sum_{i=1}^{M/2} \log\left(1 - 9\epsilon^2 \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle^2\right), \tag{4.39}$$

$$\geq \exp\left(-\sum_{t=1}^{T} \frac{18}{M}\sum_{i=1}^{M/2} \epsilon^2 \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle^2\right), \tag{4.40}$$

where in the penultimate line we have used that $O_i = -O_{i+M/2}$ along with the fact that $\log(1+x) + \log(1-x) = \log(1-x^2)$ and in the final line we have used

$\log(1-x) \geq -2x$ for all $x \in [0, 3/4]$. Now, this last line is always at least as large as the same expression maximized over all state $|\phi_{v_t}\rangle$, allowing us to write

$$\frac{p_{\text{alts}}(\ell)}{p_{\text{null}}(\ell)} \geq \exp\left(-9\epsilon^2 \sum_{t=1}^{T} \sup_{|\phi_{v_t}\rangle} \frac{2}{M} \sum_{i=1}^{M/2} \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle^2\right), \tag{4.41}$$

$$= \exp\left(-9 \cdot \epsilon^2 \cdot T \cdot \sup_{|\phi_{v_t}\rangle} \frac{2}{M} \sum_{i=1}^{M/2} \langle\phi_{v_t}| O_i |\phi_{v_t}\rangle^2\right). \tag{4.42}$$

To get a nice closed form, we will need to compute this supremum. Letting $|\phi\rangle := |\phi_{v_t}\rangle$ to ease notation, we may compute

$$\sup_{|\phi\rangle} \frac{1}{4^n - 1} \sum_{i=1}^{4^n-1} \langle\phi| P_i |\phi\rangle^2 \tag{4.43}$$

$$= \sup_{|\phi\rangle} \frac{1}{4^n - 1} \sum_{i=1}^{4^n-1} \text{tr}\left[P_i \otimes P_i |\phi\rangle\langle\phi|^{\otimes 2}\right], \tag{4.44}$$

$$= \sup_{|\phi\rangle} \frac{1}{4^n - 1} \text{tr}\left[\sum_{i=1}^{4^n-1} P_i \otimes P_i |\phi\rangle\langle\phi|^{\otimes 2}\right], \tag{4.45}$$

$$= \sup_{|\phi\rangle} \frac{1}{4^n - 1} \text{tr}\left[\left(\sum_{i=1}^{4^n-1} P_i \otimes P_i\right) |\phi\rangle\langle\phi|^{\otimes 2}\right], \tag{4.46}$$

$$= \sup_{|\phi\rangle} \frac{1}{4^n - 1} \text{tr}\left[\left(\sum_{i=1}^{4^n} P_i \otimes P_i - \mathbb{I}^{\otimes 2n}\right) |\phi\rangle\langle\phi|^{\otimes 2}\right], \tag{4.47}$$

$$= \sup_{|\phi\rangle} \frac{1}{4^n - 1} \text{tr}\left[\left(2^n \mathbb{F}_n - \mathbb{I}^{\otimes 2n}\right) |\phi\rangle\langle\phi|^{\otimes 2}\right], \quad \text{Lemma 4.2.9} \tag{4.48}$$

$$= \frac{2^n - 1}{4^n - 1}, \tag{4.49}$$

$$= \frac{1}{2^n + 1}. \tag{4.50}$$

Putting all of these pieces together, we may derive the following remarkably simple one-sided bound

$$\frac{p_{\text{alts}}(\ell)}{p_{\text{null}}(\ell)} \geq \exp\left(-\frac{9T\epsilon^2}{2^n + 1}\right) \geq 1 - \frac{9T\epsilon^2}{2^n + 1}. \tag{4.51}$$

Thus, combining Lemma 4.2.6 and Lemma 4.2.7, we may conclude

$$\Omega(1) \leq p_{\text{succ}} \leq \frac{9T\epsilon^2}{2^n + 1}. \tag{4.52}$$

For this to be true, we must take

$$T \geq \Omega\left(\frac{2^n}{\epsilon^2}\right), \tag{4.53}$$

which completes the proof. □

We utilized the following useful identity that comes up in many applications in quantum information.

Lemma 4.2.9. The SWAP operator on $(\mathbb{C}^2)^{\otimes n} \otimes (\mathbb{C}^2)^{\otimes n}$, defined by $\mathrm{SWAP} |\psi\rangle |\phi\rangle = |\phi\rangle |\psi\rangle$, admits the decomposition

$$\mathrm{SWAP} = \frac{1}{2^n} \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} P \otimes P. \tag{4.54}$$

Consequently, the operators $\{P^{\otimes 2}\}$ all commute and are simultaneously diagonalized by the n-fold Bell basis.

Exercise 4.2.1. Prove Lemma 4.2.9.

4.2.3 Pauli Shadow Tomography with Adaptive Two-copy Measurements

In the previous section we demonstrated an exponential lower bound against any single-copy measurement protocol. Our goal is to now show that given access to multi-copy, global measurements we can solve the Pauli Shadow Tomography problem much more efficiently. In fact, we can solve this using *exponentially fewer* samples if we can do measurements on $\rho^{\otimes 2}$. This is a fairly remarkable result that makes efficient Pauli shadow tomography a relatively near-term protocol.

The algorithm for learning absolute values of Paulis comes from Ref. [HKP21]. Their protocol for estimating the true value was simplified considerably in Ref. [CGY24]. Our brief treatment here follows Lec. 22 (Ch. 15) of Ref. [CC25].

Theorem 4.2.10 (Learning absolute values, [CC25])**.** There is a protocol which takes as input $O(n/\epsilon^4)$ copies of ρ, performs two-copy measurements in the Bell basis, and outputs estimates for all quantities $\{|\mathrm{tr}[P\rho]|\}$ up to additive error ϵ with high probability.

Add figure showing Bell basis measurement.

Proof. Let N denote the number of pairs of copies (i.e., we use $2N$ total copies of ρ). On each pair $\rho^{\otimes 2}$, we measure in the Bell basis. Since the Bell basis simultaneously diagonalizes all operators $P^{\otimes 2}$ for $P \in \{I, X, Y, Z\}^{\otimes n}$ (Lemma 4.2.9), each measurement outcome determines the eigenvalue of $P^{\otimes 2}$ for every Pauli P simultaneously. Denote the eigenvalue of $P^{\otimes 2}$ in the i-th measurement by $\lambda_i(P) \in \{+1, -1\}$.

Step 1: Estimating $\mathrm{tr}\,[P\rho]^2$. For each Pauli P, define the sample mean

$$\hat{\mu}_P = \frac{1}{N}\sum_{i=1}^{N} \lambda_i(P). \tag{4.55}$$

Since $\mathbb{E}[\lambda_i(P)] = \mathrm{tr}\,[P^{\otimes 2}\rho^{\otimes 2}] = \mathrm{tr}\,[P\rho]^2$ and each $\lambda_i(P) \in [-1, 1]$, Hoeffding's inequality gives

$$\Pr\Big[|\hat{\mu}_P - \mathrm{tr}\,[P\rho]^2\,| \geq t\Big] \leq 2\exp\left(-\frac{Nt^2}{2}\right) \tag{4.56}$$

for any $t > 0$. We require accuracy $t = \epsilon^2$, so

$$\Pr\Big[|\hat{\mu}_P - \mathrm{tr}\,[P\rho]^2\,| \geq \epsilon^2\Big] \leq 2\exp\left(-\frac{N\epsilon^4}{2}\right). \tag{4.57}$$

Step 2: Union bound. There are 4^n Pauli operators. By the union bound, the probability that any estimate fails is

$$\Pr\Big[\exists\, P : |\hat{\mu}_P - \mathrm{tr}\,[P\rho]^2\,| \geq \epsilon^2\Big] \leq 4^n \cdot 2\exp\left(-\frac{N\epsilon^4}{2}\right). \tag{4.58}$$

We want this to be at most δ. Setting

$$4^n \cdot 2\exp\left(-\frac{N\epsilon^4}{2}\right) \leq \delta \tag{4.59}$$

and solving for N:

$$N \geq \frac{2}{\epsilon^4}\Big(n\ln 4 + \ln(2/\delta)\Big) = O\left(\frac{n}{\epsilon^4}\right) \tag{4.60}$$

where the last equality treats δ as a constant. So with $N = O(n/\epsilon^4)$ pairs, all estimates $\hat{\mu}_P$ are simultaneously ϵ^2-accurate with probability at least $1 - \delta$.

Step 3: From squared estimates to absolute values. Output $\hat{v}_P = \sqrt{\hat{\mu}_P}$ as the estimate of $|\mathrm{tr}\,[P\rho]\,|$. (If $\hat{\mu}_P < 0$ due to statistical fluctuation, output 0.) It remains to show this is ϵ-accurate.

For any $0 \leq x, y \leq 1$:

$$(\sqrt{x} - \sqrt{y})^2 = \frac{|x-y|^2}{(\sqrt{x}+\sqrt{y})^2} = |x-y| \cdot \frac{|x-y|}{x+y+2\sqrt{xy}} \leq |x-y| \tag{4.61}$$

since the fraction is at most 1. Therefore $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$.

Setting $x = \operatorname{tr}[P\rho]^2$ and $y = \hat{\mu}_P$, and using $|x - y| \leq \epsilon^2$ from Steps 1–2:

$$\big|\,|\operatorname{tr}[P\rho]\,| - \hat{v}_P\big| = |\sqrt{x} - \sqrt{y}| \leq \sqrt{\epsilon^2} = \epsilon. \tag{4.62}$$

This holds simultaneously for all 4^n Pauli operators with probability at least $1-\delta$. □

But, what about the signs of the Paulis? Lecture 22 of Ref. [CC25] summarizes the protocol developed in Ref. [CGY24], which we will now outline in a bit more detail.

Theorem 4.2.11 (Theorem 230, [CC25]). Given ϵ-accurate estimates of $\{|\operatorname{tr}(P\rho)|\}$, and given the ability to perform single-copy measurements on $O(n/\epsilon^4)$ additional copies of ρ, there is a protocol which estimates $\{\operatorname{tr}(P\rho)\}$ to additive error $O(\epsilon)$ with high probability.

We will consider Algorithm 4 below.

Algorithm 4 LEARNPAULISIGNS$(\epsilon, \{f_P\}, \rho)$

Require: Accuracy $\epsilon > 0$, estimates $\{f_P\}$ of absolute values $\{|\operatorname{tr}(P\rho)|\}$, copies of ρ
Ensure: Estimates $\{\hat{E}_P\}$ of true values $\{\operatorname{tr}(P\rho)\}$

1: Find a state σ such that $|f_P - |\operatorname{tr}(P\sigma)|| \leq \epsilon$ for every P.
2: Perform Bell measurements on $O(n/\epsilon^4)$ copies of $\sigma \otimes \rho$ to estimate all $\operatorname{tr}(P^{\otimes 2}(\sigma \otimes \rho)) = \operatorname{tr}(P\sigma)\operatorname{tr}(P\rho)$ to error ϵ^2 with high probability.
3: Denote each estimate of $\operatorname{tr}(P\sigma)\operatorname{tr}(P\rho)$ by g_P.
4: If $f_P < 2\epsilon$, set $\hat{E}_P = 0$. Otherwise if $f_P > 2\epsilon$, set $\hat{E}_P = g_P/\operatorname{tr}(\sigma P)$.
5: **return** $\{\hat{E}_P\}$

Proof. The existence of a state σ satisfying $|f_P - |\operatorname{tr}(P\rho)|| \leq \epsilon$ for all P, should be clear as ρ itself satisfies the condition. However, we note that actually finding this state may be computationally inefficient. Here, though, we are only concerned with sample efficiency. Once we have an explicit description of this σ, we get exact access[2] to $\operatorname{tr}(P\sigma)$.

Now, if $f_P < 2\epsilon$, we set $\hat{E}_P = 0$ and this is trivially a 3ϵ-accurate estimate of $\operatorname{tr}(P\rho)$:

$$|\hat{E}_P - \operatorname{tr}(P\rho)| = |\operatorname{tr}(P\rho) - f_P + f_P| \leq |\operatorname{tr}(P\rho) - f_P| + |f_P| \leq \epsilon + 2\epsilon = 3\epsilon. \tag{4.63}$$

[2] Again, for large system sizes and without any additional structure, the classical memory will be a bottleneck here.

If $f_P > 2\epsilon$, we set $\hat{E}_P = g_P / \operatorname{tr}(P\sigma)$, in which case

$$|\hat{E}_P - \operatorname{tr}(P\rho)| = \frac{|g_P - \operatorname{tr}(P\sigma)\operatorname{tr}(P\rho)|}{|\operatorname{tr}(P\sigma)|} \leq \frac{\epsilon^2}{f_P - \epsilon} \leq \epsilon. \tag{4.64}$$

□

4.2.4 Project Idea: Adaptivity Can Help Exponentially in Shadow Tomography

The adaptivity in Algorithm 4—first estimating absolute values, then using those results to choose a measurement strategy for the signs—is not merely a convenience but a genuine necessity. Ref. [CGZ24] proved that any *nonadaptive* two-copy protocol for estimating $\operatorname{tr}(P\rho)$ for all Paulis P requires $\Omega(2^{n/2})$ copies of ρ, while the adaptive two-round protocol of [CGY24] achieves this with only $O(n)$ copies (at constant ϵ). This is an exponential separation—the first known for any quantum learning task—and it arises precisely because resolving the signs requires a measurement basis that depends on the outcomes of the first round. The nonadaptive lower bound follows from a Le Cam-style argument: if the measurements are fixed in advance, two-copy measurements cannot distinguish states $(I + P)/2^n$ from $(I - P)/2^n$ (which differ only in the sign of a single Pauli expectation value) without exponentially many samples. Adaptivity breaks this barrier by allowing the second-round measurements to be tailored to the specific state, informed by the absolute values learned in the first round.

5 Quantum Property Testing

> *Pure or far from pure, that is the question.*
>
> **— Shakespeare?**

5.1 Essential Definitions

We now turn to the topic of *quantum property testing*. See Ref. [MD16] for a great survey of the field as it was in 2016.

Definition 5.1.1 (Property Testing). Let $\mathcal{X}$ be a set of objects and $d : \mathcal{X} \times \mathcal{X} \to [0, 1]$ be a distance measure on $\mathcal{X}$. A *property* is any subset $\mathcal{P} \subseteq \mathcal{X}$. Let $x \in \mathcal{X}$. Then, if $\inf_{y \in \mathcal{P}} d(x, y) \geq \epsilon$, we say x is ϵ-far from $\mathcal{P}$. Otherwise, we say it is ϵ-close.

An ϵ-*tester* for $\mathcal{P}$ is an algorithm that takes as input either an $x \in \mathcal{P}$ or an x that is ϵ-far from $\mathcal{P}$. Conventionally, one requires:

- **Completeness.** $\Pr[\text{Accept}|x \in \mathcal{P}] \geq 2/3$,
- **Soundness.** $\Pr[\text{Accept}|x\ \epsilon\text{−far}] \leq 1/3$.

The difference between the completeness and soundness is called the *bias* of the algorithm, which we require to be at least a constant.

Quick Quiz 5.1.2. Our definition of an ϵ-tester requires completeness 2/3 and soundness 1/3. Is this convention without loss of generality? In other words, would any constants work or is there something special about these?

Clearly if we accepted either case with equal probability, this would be an utterly useless tester; we learn nothing by running the algorithm. Perturbing slightly away from this vacuous edge case, suppose we have a completeness $c = 1/2 + \delta/2$ and a soundness $s = 1/2 + \delta/2$ (i.e. we have a bias of $\delta > 0$). We can boost this non-zero

bias to any other desired constant by repeating the algorithm R times and taking a majority vote (i.e. accepting only if at least $R/2$ trials accept). How many repetitions R suffice?

Theorem 5.1.3. Let $\mathcal{A}$ be an ϵ-tester for a property $\mathcal{P}$ with completeness $\frac{1}{2}+\frac{b}{2}$ and soundness $\frac{1}{2}-\frac{b}{2}$, where $b>0$ is the bias. Then for any $\eta \in (0,1/2)$, there exists an ϵ-tester $\mathcal{A}'$ with completeness $1-\eta$ and soundness η that uses $O\left(\frac{q}{b^2}\ln\frac{1}{\eta}\right)$ queries, where q is the query/sample complexity of $\mathcal{A}$.

Proof. Run $\mathcal{A}$ independently R times (to be chosen) and take a majority vote. Let $X_i \in \{0,1\}$ indicate that the i-th run accepts, and let $S=\sum_{i=1}^{R} X_i$. Each X_i is Bernoulli with parameter $p=\Pr[\mathcal{A} \text{ accepts } x]$.

Completeness ($x\in\mathcal{P}$, so $p\geq\frac{1}{2}+\frac{b}{2}$). The amplified tester errs if $S\leq R/2$. Since $\mathbb{E}[S]=pR\geq(\frac{1}{2}+\frac{b}{2})R$, this requires S to fall at least $\frac{b}{2}R$ below its mean:

$$\Pr\left[S\leq\frac{R}{2}\right] \;\leq\; \Pr\left[S\leq pR-\frac{b}{2}R\right] \;\leq\; \exp\left(-\frac{b^2R}{2}\right),$$

where the last step is Hoeffding's inequality.

Soundness (x is ϵ-far from $\mathcal{P}$, so $p\leq\frac{1}{2}-\frac{b}{2}$). The amplified tester errs if $S>R/2$. Since $\mathbb{E}[S]=pR\leq(\frac{1}{2}-\frac{b}{2})R$, this requires S to exceed its mean by at least $\frac{b}{2}R$:

$$\Pr\left[S>\frac{R}{2}\right] \;\leq\; \Pr\left[S\geq pR+\frac{b}{2}R\right] \;\leq\; \exp\left(-\frac{b^2R}{2}\right).$$

Choosing R. Setting $\exp(-b^2R/2)\leq\eta$ and solving:

$$\boxed{R \;=\; \left\lceil\frac{2\ln(1/\eta)}{b^2}\right\rceil \;=\; O\left(\frac{1}{b^2}\right)}$$

where the $O(\cdot)$ treats η as a fixed constant. Thus, if an algorithm has sample complexity q to achieve a bias b, we can boost the bias to any fixed constant η using $R\cdot q$ total samples. □

With this definition in place, let us considering perhaps the simplest property testing task.

5.1.1 Equality to Fixed Pure State

For starters let us consider the property of being equal to a given pure state ϕ. Formally, EQUALITY TO can be expressed as

$$\mathcal{P} = \{e^{i\theta} |\phi\rangle : \theta \in \mathbb{R}\}. \tag{5.1}$$

Note that the phase freedom is essential because no measurement can distinguish between $|\phi\rangle$ and $e^{i\theta} |\phi\rangle$.

> **Quick Quiz 5.1.4.** Without worrying about experimental feasibility, what is the most natural way to test equality to a fixed pure state?

The simplest tester performs the two-outcome projective measurement $\{|\phi\rangle\langle\phi|,\ I - |\phi\rangle\langle\phi|\}$ on a single copy of $|\psi\rangle$, and accepts if and only if the first outcome is obtained. Given an unknown state $|\psi\rangle$, the probability that the first outcome is observed is given as

$$\Pr[\text{Accept}] = |\langle\psi|\phi\rangle|^2 . \tag{5.2}$$

We may then compute the completeness and soundness as follows.

Completeness. If $|\psi\rangle \in \mathcal{P}$, then $|\psi\rangle = e^{i\theta} |\phi\rangle$, so

$$\Pr[\text{Accept}] \ = \ |\langle\psi|\phi\rangle|^2 = |\langle\psi|\psi\rangle|^2 \ = \ 1.$$

Soundness. If $D(|\psi\rangle , |\phi\rangle) = \epsilon$, then $|\langle\psi|\phi\rangle|^2 = 1 - \epsilon^2$, so

$$\Pr[\text{Reject}] \ = \ 1 - |\langle\psi|\phi\rangle|^2 \ = \ \epsilon^2.$$

A single copy thus gives completeness 1 and soundness $1 - \epsilon^2$. The bias is ϵ^2, which may be small.

Sample complexity. By the amplification lemma, repeating $R = O(1/\epsilon^4)$ times and taking a majority vote boosts to constant completeness and soundness. But we can do better: since completeness is *perfect* (i.e., 1), we can instead accept if and only if *all* R runs accept. Completeness remains 1. In the soundness case, each run accepts with probability $1 - \epsilon^2$, so

$$\Pr[\text{all } R \text{ runs accept} : \epsilon\text{-far}] \ = \ (1 - \epsilon^2)^R \ \leq \ e^{-\epsilon^2 R}.$$

Setting $e^{-\epsilon^2 R} \leq 1/3$ gives $R = O(1/\epsilon^2)$.

Theorem 5.1.5 (General lower bound for pure state properties)**.** Let $\mathcal{P}$ be any non-trivial property of pure states (i.e., there exist $|\phi\rangle \in \mathcal{P}$ and $|\phi'\rangle \notin \mathcal{P}$ with $D(|\phi\rangle, |\phi'\rangle) = \varepsilon$). Then any ε-tester for $\mathcal{P}$ with success probability at least $2/3$ requires $\Omega(1/\varepsilon^2)$ copies.

Proof. Let $|\phi\rangle \in \mathcal{P}$ and $|\phi'\rangle \notin \mathcal{P}$ with $D(|\phi\rangle, |\phi'\rangle) = \varepsilon$ be the promised pair. Consider the input drawn uniformly from $\{|\phi\rangle, |\phi'\rangle\}$. Any ε-tester using k copies that accepts $|\phi\rangle$ with probability $\geq 2/3$ and rejects $|\phi'\rangle$ with probability $\geq 2/3$ must, in particular, be able to distinguish $|\phi\rangle\langle\phi|^{\otimes k}$ from $|\phi'\rangle\langle\phi'|^{\otimes k}$ with success probability $\geq 2/3$.

By the Holevo-Helstrom theorem (Thm. 2.2.9), the optimal success probability for this binary discrimination task is

$$p_{\text{succ}} = \frac{1 + D_{\text{tr}}\left(|\phi\rangle\langle\phi|^{\otimes k}, |\phi'\rangle\langle\phi'|^{\otimes k}\right)}{2}.$$

For pure states, the trace distance evaluates to

$$D_{\text{tr}}\left(|\phi\rangle\langle\phi|^{\otimes k}, |\phi'\rangle\langle\phi'|^{\otimes k}\right) = \sqrt{1 - |\langle\phi|\phi'\rangle|^{2k}} = \sqrt{1 - (1-\varepsilon^2)^k}.$$

Requiring $p_{\text{succ}} \geq 2/3$ forces $D_{\text{tr}} \geq 1/3$, i.e., $(1-\varepsilon^2)^k \leq 8/9$. Since $\ln(1-\varepsilon^2) \leq -\varepsilon^2$, this gives

$$k \geq \frac{\ln(9/8)}{\varepsilon^2} = \Omega(1/\varepsilon^2). \qquad \square$$

Corollary 5.1.6 (Lower bound for Equality to $|\phi\rangle$)**.** Any ε-tester for **Equality to** $|\phi\rangle$ with success probability at least $2/3$ requires $\Omega(1/\varepsilon^2)$ copies.

Quick Quiz 5.1.7. What if, instead of testing equality to a known pure state, we are asked to test whether two unknown pure states are close to one another or at least ϵ-far?

We might call this the EQUALITY property. To construct a natural algorithm for this task, we must introduce an important primitive called the SWAP test.

5.2 The SWAP Test

The circuit in Fig. 5.1 was first introduced in Ref. [Bar+97] in the context of stabilizing quantum computations. A few years after this paper, Refs. [Buh+01; Eke+02] used the same circuit for very similar tasks. It has since been come to be

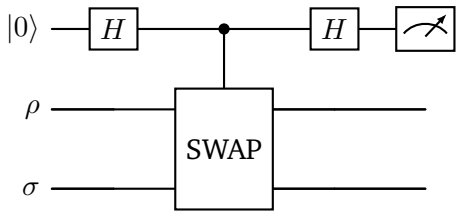

Fig. 5.1: The SWAP Test.

called *the SWAP test* and is a ubiquitous primitive in quantum information [Nis25].

Exercise 5.2.1. Show that the probability of obtaining outcome "0" when measuring the ancilla qubit in the computational basis is given as

$$p(0) = \frac{1}{2} + \frac{1}{2}\mathrm{tr}\left[\rho\sigma\right]. \tag{5.3}$$

5.3 Testing Equality between Pure States

Exercise 5.3.1. Show that the SWAP test can be used to construct a tester for EQUALITY using at most $O(1/\epsilon^2)$ samples.

5.4 Purity Testing

Let us see the first example of testing a property of mixed quantum states. The PURITY property is formally defined as

$$\mathcal{P} = \{\rho \in \mathcal{B}(\mathbb{C}^d) : \mathrm{tr}\left[\rho^2\right] = 1\}. \tag{5.4}$$

5.4.1 Sample-efficient Multi-copy Algorithm via the SWAP Test

Proposition 5.4.1 (Purity testing via the swap test)**.** The property PURITY—whether an unknown n-qubit state ρ is pure—can be tested with $O(1/\epsilon)$ copies of ρ.

Proof. The swap test on two copies of ρ accepts (i.e. outputs "pure") with probability

$$p_{\text{acc}} = \frac{1 + \operatorname{tr}(\rho^2)}{2}. \tag{5.5}$$

Completeness. If ρ is pure, then $\operatorname{tr}(\rho^2) = 1$, so $p_{\text{acc}} = 1$. The test accepts with certainty; in particular, repeating it any number of times always accepts. Thus the protocol has *perfect completeness*.

Soundness. Suppose ρ is ϵ-far from every pure state in trace distance. We claim that $\operatorname{tr}(\rho^2) \leq 1 - \epsilon$, which implies $p_{\text{acc}} \leq 1 - \epsilon/2$.

To see this, let $\rho = \sum_i \lambda_i |\psi_i\rangle\langle\psi_i|$ be the eigendecomposition with eigenvalues in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. The closest pure state to ρ is $|\psi_1\rangle$, and the assumption that ρ is ϵ-far from pure means $\lambda_1 \leq 1 - \epsilon$. Now observe that

$$\operatorname{tr}(\rho^2) = \sum_i \lambda_i^2 \leq \max_i \lambda_i \cdot \sum_j \lambda_j = \lambda_1 \cdot 1 = \lambda_1 \leq 1 - \epsilon, \tag{5.6}$$

where we used $\sum_j \lambda_j = \operatorname{tr}(\rho) = 1$. Substituting into (5.5):

$$p_{\text{acc}} = \frac{1 + \operatorname{tr}(\rho^2)}{2} \leq \frac{1 + (1 - \epsilon)}{2} = 1 - \frac{\epsilon}{2}. \tag{5.7}$$

Sample complexity via repetition. We run T independent swap tests, each consuming two copies of ρ, and accept if and only if all T tests accept. Let $X_i \in \{0, 1\}$ be the indicator for the i-th test accepting, and let $S = \sum_{i=1}^{T} X_i$.

If ρ is pure, all tests accept with certainty, so $\Pr[S = T] = 1$.

If ρ is ϵ-far from pure, each test accepts independently with probability at most $1 - \epsilon/2$, so

$$\Pr[S = T] = p_{\text{acc}}^T \leq \left(1 - \frac{\epsilon}{2}\right)^T \leq e^{-\epsilon T/2}. \tag{5.8}$$

Setting $e^{-\epsilon T/2} \leq 1/3$ (or any desired failure probability δ) gives

$$T \geq \frac{2 \log 3}{\epsilon} = O(1/\epsilon). \tag{5.9}$$

Since each swap test uses two copies, the total number of copies is $2T = O(1/\epsilon)$. □

5.4.2 Sample Complexity Lower Bound

Proposition 5.4.2 (Lower bound for purity testing)**.** Testing whether a mixed state is pure or ϵ-far (in trace distance) requires $T = \Omega(1/\epsilon)$ copies of ρ.

Proof. For $\epsilon \in (0, 1)$, consider the problem of distinguishing two single-qubit mixed states ρ_0 and ρ_ϵ, where

$$\rho_x = (1 - x)|0\rangle\langle 0| + x|1\rangle\langle 1|. \tag{5.10}$$

Note that $\rho_0 = |0\rangle\langle 0|$ is a pure state, and ρ_ϵ is ϵ-far from ρ_0 in trace distance. Any purity tester with sample complexity T can be used to distinguish $\rho_0^{\otimes T}$ from $\rho_\epsilon^{\otimes T}$ with success probability $p_{\text{succ}} \geq 2/3$.

By the Holevo–Helstrom theorem, the optimal success probability for distinguishing two states σ_0, σ_1 with equal prior is

$$p_{\text{succ}} \leq \frac{1}{2}(1 + d_{\text{tr}}(\sigma_0, \sigma_1)), \tag{5.11}$$

where $d_{\text{tr}}(\cdot, \cdot)$ denotes the trace distance. Applying this to $\sigma_0 = \rho_0^{\otimes T}$ and $\sigma_1 = \rho_\epsilon^{\otimes T}$, we bound the trace distance via the Fuchs–van de Graaf inequality:

$$d_{\text{tr}}(\rho_0^{\otimes T}, \rho_\epsilon^{\otimes T}) \leq \sqrt{1 - \mathrm{F}(\rho_0^{\otimes T}, \rho_\epsilon^{\otimes T})^2}, \tag{5.12}$$

where $\mathrm{F}(\rho, \sigma) = \mathrm{tr}[\sqrt{\sqrt{\sigma}\rho\sqrt{\sigma}}]$ is the fidelity. Since the fidelity is multiplicative under tensor products,

$$\mathrm{F}(\rho_0^{\otimes T}, \rho_\epsilon^{\otimes T})^2 = \mathrm{F}(\rho_0, \rho_\epsilon)^{2T}. \tag{5.13}$$

Now, since $\rho_0 = |0\rangle\langle 0|$ is pure,

$$\mathrm{F}(\rho_0, \rho_\epsilon)^2 = \langle 0|\rho_\epsilon|0\rangle = 1 - \epsilon. \tag{5.14}$$

Combining, we obtain

$$p_{\text{succ}} \leq \frac{1}{2}\Big(1 + \sqrt{1 - (1 - \epsilon)^T}\Big). \tag{5.15}$$

Imposing $p_{\text{succ}} \geq 2/3$ and rearranging:

$$\sqrt{1 - (1 - \epsilon)^T} \geq \frac{1}{3} \quad \Longrightarrow \quad (1 - \epsilon)^T \leq \frac{8}{9}. \tag{5.16}$$

Taking logarithms and using $\log(1 - \epsilon) \leq -\epsilon$:

$$S \geq \frac{\log(9/8)}{-\log(1 - \epsilon)} \geq \frac{\log(9/8)}{\epsilon} = \Omega(1/\epsilon). \tag{5.17}$$

□

6 Solutions to Exercises

> *Mathematics, you see, is not a spectator sport. To understand mathematics means to be able to do mathematics. And what does it mean [to be] doing mathematics? In the first place, it means to be able to solve mathematical problems.*
>
> **— George Pólya**

In this appendix, we will provide the solutions to the exercises that appear at the end of each section.

Solution to Exercise 2.1.1. See Lecture 2 of Ref. [Wri24].

Solution to Exercise 2.1.2. See Lecture 1 and 2 of Ref. [Wri24].

Solution to Exercise 2.2.1.

Proof. Let the two states be given with equal priors $p_0 = p_1 = \frac{1}{2}$. Using Theorem 2.2.9, we know that

$$p_{\text{succ}} = \frac{1}{2} + \frac{1}{2} d_{\text{tr}}(|\psi\rangle, |\phi\rangle).$$

To compute the trace distance, we invoke the **Fuchs-van de Graaf** relation. For pure states, the inequality saturates to an equality:

$$d_{\text{tr}}(|\psi\rangle, |\phi\rangle) = \sqrt{1 - F(|\psi\rangle, |\phi\rangle)}$$

where the fidelity F for pure states is the squared overlap:

$$F(|\psi\rangle, |\phi\rangle) = |\langle\psi|\phi\rangle|^2$$

Given that the states differ by an angle θ, we have $|\langle\psi|\phi\rangle| = \cos(\theta)$. Substituting this into the fidelity:

$$F = \cos^2(\theta)$$

Now, substituting F back into the trace distance expression:

$$D(|\psi\rangle, |\phi\rangle) = \sqrt{1 - \cos^2(\theta)} = \sqrt{\sin^2(\theta)} = \sin(\theta)$$

(assuming $\theta \in [0, \pi/2]$).

Finally, substituting $d_{\text{tr}} = \sin(\theta)$ back into the Holevo-Helstrom equation:

$$p_{\text{succ}} = \frac{1}{2} + \frac{1}{2}\sin(\theta)$$

□

Solution to Exercise 2.2.2

Proof. Because A, B are Hermitian, we can diagonalize them in terms of some orthonormal basis $\{|u_i\rangle\}$ and $\{|v_i\rangle\}$ such that

$$A = \sum_{i=1}^{d} p_i |u_i\rangle \langle u_i| \quad \text{and} \quad B = \sum_{j=1}^{d} q_j |v_j\rangle \langle v_j| . \tag{6.1}$$

This allows us to write

$$\begin{aligned}
\mathrm{tr}\left[AB\right] &= \mathrm{tr}\left[\sum_i p_i \left|u_i\right\rangle\left\langle u_i\right| \sum_j q_j \left|v_j\right\rangle\left\langle v_j\right|\right], && (6.2)\\
&= \sum_{ij} p_i q_j \mathrm{tr}\left[\left|u_i\right\rangle\left\langle u_i|v_j\right\rangle\left\langle v_j\right|\right], && (6.3)\\
&= \sum_{ij} p_i q_j \left|\left\langle u_i|v_j\right\rangle\right|^2, && (6.4)\\
&= \sum_{ij} p_i q_j \left(\left|\left\langle u_i|v_j\right\rangle\right|^2\right)^1, && (6.5)\\
&= \sum_{ij} p_i q_j \left(\left|\left\langle u_i|v_j\right\rangle\right|^2\right)^{\frac{1}{p}+\frac{1}{q}}, && (6.6)\\
&= \sum_{ij} \left(p_i \left|\left\langle u_i|v_j\right\rangle\right|^{\frac{2}{p}}\right)\left(q_j \left|\left\langle u_i|v_j\right\rangle\right|^{\frac{2}{q}}\right), && (6.7)\\
&\leq \left(\sum_{ij} p_i^p \left|\left\langle u_i|v_j\right\rangle\right|^2\right)^{\frac{1}{p}} \left(\sum_{ij} q_j^q \left|\left\langle u_i|v_j\right\rangle\right|^2\right)^{\frac{1}{q}}, \quad \text{Hölder's Inequality} && (6.8)\\
&= \left(\sum_i p_i^p \sum_j \left|\left\langle u_i|v_j\right\rangle\right|^2\right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i \left|\left\langle u_i|v_j\right\rangle\right|^2\right)^{\frac{1}{q}}, && (6.9)\\
&= \left(\sum_i p_i^p \sum_j \left\langle u_i|v_j\right\rangle\left\langle v_j|u_i\right\rangle\right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i \left\langle v_j|u_i\right\rangle\left\langle u_i|v_j\right\rangle\right)^{\frac{1}{q}}, && (6.10)\\
&= \left(\sum_i p_i^p\right)^{\frac{1}{p}} \left(\sum_j q_j^q\right)^{\frac{1}{q}}, && (6.11)\\
&= \|A\|_p \|B\|_q, && (6.12)
\end{aligned}$$

where to obtain the penultimate line, we used the fact that $\{|u_i\rangle\}$ and $\{|v_i\rangle\}$ form orthonormal bases, thus allowing us to resolve the identity. □

Solution to Exercise 2.2.3

Second Proof of Holevo-Helstrom. Suppose we have a two-element POVM $\{E_1, E_2\}$ such that $E_1 + E_2 = \mathbb{I}$. Our algorithm will be to simply return ρ_i when outcome i is

observed. Assuming the probability of having ρ_1 and ρ_2 is the same, we can express the success probability as

$$\begin{aligned} p_{\text{succ}} &= \frac{1}{2}\text{tr}\left[E_1\rho_1\right] + \frac{1}{2}\text{tr}\left[\rho_2 E_2\right], && (6.13)\\ &= \frac{1}{2}\text{tr}\left[E_1\rho_1 + \rho_2 E_2\right], && (6.14)\\ &= \frac{1}{4}\text{tr}\left[E_1\rho_1 + \rho_2 E_2\right] + \frac{1}{4}\text{tr}\left[E_1\rho_1 + \rho_2 E_2\right], && (6.15)\\ &= \frac{1}{4}\text{tr}\left[(E_1 + E_2)(\rho_1 + \rho_2)\right] + \frac{1}{4}\text{tr}\left[(E_1 - E_2)(\rho_1 - \rho_2)\right], && (6.16)\\ &= \frac{1}{2} + \frac{1}{4}\text{tr}\left[(E_1 - E_2)(\rho_1 - \rho_2)\right], && (6.17)\\ &=: \frac{1}{2} + \frac{1}{2}T, && (6.18)\\ & && (6.19) \end{aligned}$$

where we have used that $E_1 + E_2 = \mathbb{I}$ and $\text{tr}\left[\rho_i\right] = 1$. Now, we can apply Holder's inequality

$$\begin{aligned} T &= \frac{1}{2}\text{tr}\left[(E_1 - E_2)(\rho_1 - \rho_2)\right], && (6.20)\\ &\leq \frac{1}{2}\|E_1 - E_2\|_\infty \|\rho_1 - \rho_2\|_1, && (6.21)\\ &\leq \frac{1}{2}\|\rho_1 - \rho_2\|_1, && (6.22)\\ &= d_{\text{tr}}(\rho_1, \rho_2) && (6.23) \end{aligned}$$

where we have used that the maximum eigenvalue of $\rho_1 - \rho_2$ is less than unity because POVM elements are, by definition, between 0 and $\mathbb{I}$. Putting these together, we have

$$p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2}d_{\text{tr}}(\rho_1, \rho_2), \tag{6.24}$$

as desired. We won't show the optimal measurement strategy, because it is the same as above. □

Solution to Exercise 3.3.1. Recall that the dimension of a finite dimensional vector space can be computed via the trace of an orthogonal projector onto that subspace. Thus, we may determine the dimension of the symmetric subspace by computing

$$\text{tr}\left[\Pi_{\text{sym}}^{d,n}\right] = \dim \vee^n \mathbb{C}^d = \binom{n+d-1}{n}. \tag{6.25}$$

To this end, we write

$$\operatorname{tr}\left[\Pi_{\text{sym}}^{d,n}\right] = \frac{1}{n!}\sum_{\pi\in S_n}\operatorname{tr}\left[P_d(\pi)\right]. \tag{6.26}$$

Thus, we first must determine the trace of an arbitrary permutation matrix. To do this, first let $c(\pi)$ count the number of cycles in π. For example, let $\pi = (134)(2)(5)$ be a permutation in S_5. In this case, $c(\pi) = 3$ because there are 3 disjoint cycles. With this definition in place, observe

$$\operatorname{tr}\left[P_d(\pi)\right] = \sum_{i_1,\ldots,i_n\in[d]}\prod_{j=1}^{n}\langle i_j|i_{\pi^{-1}(j)}\rangle. \tag{6.27}$$

This product yields a 1 if and only if $i_j = i_{\pi^{-1}(j)}$ for all $j \in [n]$. To be invariant under π, a basis vector $|i_1, i_2, \ldots, i_n\rangle$ must satisfy

$$i_k = i_{\pi(k)} \quad \text{for all } k. \tag{6.28}$$

This means that the values i_k must be constant on each cycle of π. Since the cycles of π are disjoint, the constraints are independent. We can reason as follows:

- Each disjoint cycle of π imposes a constraint that the indices involved in the cycle must be equal.
- For each cycle, we can choose any value from $\{1, \ldots, d\}$ to assign to all positions in that cycle.
- There are $c(\pi)$ disjoint cycles.

Therefore, the number of basis vectors fixed by $P_d(\pi)$ is: $d^{c(\pi)}$. Each of these contributes 1 to the sum, yielding

$$\operatorname{tr}\left[P_d(\pi)\right] = \sum_{i_1,\ldots,i_n\in[d]}\prod_{j=1}^{n}\langle i_j|i_{\pi^{-1}(j)}\rangle = d^{c(\pi)}. \tag{6.29}$$

Now, returning to our original goal, we must compute the sum

$$\operatorname{tr}\left[\Pi_{\text{sym}}^{d,n}\right] = \frac{1}{n!}\sum_{\pi\in S_n} d^{c(\pi)}. \tag{6.30}$$

We claim that this is equal to the dimension of the symmetric subspace. The claim follows from the proof of the following lemma.

Lemma 6.0.1 (Rising factorial, Pochhammer Symbol)**.** Let d, n be integers and let $c(\pi)$ count the number of disjoint cycles comprising $\pi \in S_n$. Then, the following holds

$$\sum_{\pi \in S_n} d^{c(\pi)} = d(d+1)\cdots(d+n+1) = \binom{n+d-1}{n} n! \tag{6.31}$$

Proof. By induction on n. The base case $n = 1$ gives $d = d$. For the inductive step, any $\pi \in S_{n+1}$ is obtained from some $\pi' \in S_n$ in exactly two ways: either $n+1$ is a fixed point, or it is inserted immediately after one of the n elements of an existing cycle. In the first case, $c(\pi) = c(\pi') + 1$, so each π' contributes $d^{c(\pi')+1} = d \cdot d^{c(\pi')}$. In the second case, $c(\pi) = c(\pi')$, but each π' yields n distinct permutations each with weight $d^{c(\pi')}$, contributing $n \cdot d^{c(\pi')}$. Summing over all $\pi' \in S_n$ gives the recursion

$$f(n+1) = (d+n) f(n),$$

so $f(n) = d(d+1)\cdots(d+n-1) = \frac{(d+n-1)!}{(d-1)!} = \binom{n+d-1}{n} n!$. □

Solution to Exercise 3.6.1. Write $p = 1/2 + \delta$ so that $|p - 1/2| = |\delta|$. We use a second-order Taylor expansion of H around $p = 1/2$. Computing the first two derivatives:

$$H'(p) = \log \frac{1-p}{p}, \qquad H''(p) = -\frac{1}{p(1-p)}. \tag{6.32}$$

Evaluating at $p = 1/2$ gives $H(1/2) = 1$ and $H'(1/2) = 0$. For the second derivative, we note that for all $p \in [0, 1]$:

$$p(1-p) \leq \frac{1}{4}, \tag{6.33}$$

and therefore

$$H''(p) = -\frac{1}{p(1-p)} \geq -4. \tag{6.34}$$

By Taylor's theorem with remainder, there exists ξ between $1/2$ and p such that:

$$H(p) = H(1/2) + H'(1/2)\delta + \frac{1}{2} H''(\xi)\delta^2 = 1 + \frac{1}{2} H''(\xi)\delta^2. \tag{6.35}$$

Applying the lower bound $H''(\xi) \geq -4$:

$$H(p) \geq 1 + \frac{1}{2}(-4)\delta^2 = 1 - 2\delta^2 \geq 1 - 4\delta^2, \tag{6.36}$$

where the last inequality holds since $\delta^2 \geq 0$. This completes the proof.

Solution to Exercise 4.2.1. Recall, our goal is to show that the SWAP operator on $(\mathbb{C}^2)^{\otimes n} \otimes (\mathbb{C}^2)^{\otimes n}$, defined by $\mathrm{SWAP}\,|\psi\rangle\,|\phi\rangle = |\phi\rangle\,|\psi\rangle$, admits the decomposition

$$\mathrm{SWAP} = \frac{1}{2^n} \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} P \otimes P. \tag{6.37}$$

Consequently, the operators $\{P^{\otimes 2}\}$ all commute and are simultaneously diagonalized by the n-fold Bell basis.

Proof. We proceed in three steps.

Step 1: The $n = 1$ case. We show that

$$\mathrm{SWAP}_1 = \frac{1}{2}(I \otimes I + X \otimes X + Y \otimes Y + Z \otimes Z) \tag{6.38}$$

on $\mathbb{C}^2 \otimes \mathbb{C}^2$. The set $\{I, X, Y, Z\}$ forms an orthogonal basis for the space of 2×2 matrices under the Hilbert–Schmidt inner product $\langle A, B\rangle = \mathrm{tr}\left[A^\dagger B\right]$, with $\mathrm{tr}\left[P^\dagger P\right] = 2$ for each Pauli P. Any operator M on $\mathbb{C}^2 \otimes \mathbb{C}^2$ can therefore be expanded as

$$M = \frac{1}{4} \sum_{\sigma,\tau \in \{I,X,Y,Z\}} \mathrm{tr}\left[(\sigma \otimes \tau)^\dagger M\right]\, \sigma \otimes \tau. \tag{6.39}$$

For $M = \mathrm{SWAP}$, the expansion coefficients are

$$\mathrm{tr}\left[(\sigma \otimes \tau)^\dagger \mathrm{SWAP}\right] = \mathrm{tr}\left[\sigma\tau^\dagger\right] \tag{6.40}$$

where the right-hand side follows from the action of SWAP, which contracts the two tensor factors. By orthogonality of the Paulis, $\mathrm{tr}\left[\sigma\tau^\dagger\right] = 2\,\delta_{\sigma\tau}$, so only the $\sigma = \tau$ terms survive:

$$\mathrm{SWAP}_1 = \frac{1}{4}\sum_{\sigma} \mathrm{tr}\left[\sigma\sigma^\dagger\right]\, \sigma \otimes \sigma = \frac{1}{4}\sum_{\sigma} 2 \cdot \sigma \otimes \sigma = \frac{1}{2} \sum_{\sigma \in \{I,X,Y,Z\}} \sigma \otimes \sigma. \tag{6.41}$$

Step 2: Extension to n qubits. The n-qubit SWAP factors as a tensor product of single-qubit SWAPs acting on corresponding pairs of qubits:

$$\mathrm{SWAP}_n = \mathrm{SWAP}_1^{(1)} \otimes \mathrm{SWAP}_1^{(2)} \otimes \cdots \otimes \mathrm{SWAP}_1^{(n)} \tag{6.42}$$

where $\mathrm{SWAP}_1^{(k)}$ swaps the k-th qubit of the first register with the k-th qubit of the second. Substituting Step 1 into each factor:

$$\mathrm{SWAP}_n = \prod_{k=1}^{n} \left(\frac{1}{2} \sum_{\sigma_k \in \{I,X,Y,Z\}} \sigma_k \otimes \sigma_k \right). \tag{6.43}$$

Expanding the product distributes over the sums, giving

$$\mathrm{SWAP}_n = \frac{1}{2^n} \sum_{\sigma_1,\ldots,\sigma_n} (\sigma_1 \otimes \cdots \otimes \sigma_n) \otimes (\sigma_1 \otimes \cdots \otimes \sigma_n) = \frac{1}{2^n} \sum_{P \in \{I,X,Y,Z\}^{\otimes n}} P \otimes P. \tag{6.44}$$

Step 3: Simultaneous diagonalizability. First, every $P^{\otimes 2}$ commutes with every $Q^{\otimes 2}$: since P and Q are Pauli operators, they satisfy $PQ = cQP$ for some phase $c \in \{+1, -1\}$ (i.e., they either commute or anticommute). Then

$$P^{\otimes 2}Q^{\otimes 2} = (PQ)^{\otimes 2} = (cQP)^{\otimes 2} = c^2(QP)^{\otimes 2} = (QP)^{\otimes 2} = Q^{\otimes 2}P^{\otimes 2} \tag{6.45}$$

where the phase squares to 1 regardless of its sign. As a commuting family of Hermitian operators, the $\{P^{\otimes 2}\}$ are simultaneously diagonalizable.

It remains to identify the common eigenbasis. Since SWAP_1 is Hermitian with eigenvalues ± 1, its eigenspaces are the symmetric and antisymmetric subspaces of $\mathbb{C}^2 \otimes \mathbb{C}^2$, which are spanned by the Bell states. More precisely, the four Bell states

$$|\Phi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle), \qquad |\Psi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle) \tag{6.46}$$

are simultaneous eigenstates of $\sigma \otimes \sigma$ for every $\sigma \in \{I, X, Y, Z\}$. Since every n-qubit operator $P^{\otimes 2}$ is a tensor product of such single-qubit factors $\sigma_k \otimes \sigma_k$, the n-fold Bell basis — consisting of tensor products of Bell states across each pair of corresponding qubits — simultaneously diagonalizes all $P^{\otimes 2}$. □

Bibliography

[Ach+25a] Jayadev Acharya, Abhilash Dharmavarapu, Yuhan Liu, and Nengkun Yu. *Pauli Measurements Are Near-Optimal for Single-Qubit Tomography*. July 2025. arXiv: 2507.22001 [quant-ph] (cit. on p. 34).

[Ach+25b] Jayadev Acharya, Abhilash Dharmavarapu, Yuhan Liu, and Nengkun Yu. *Pauli Measurements Are Not Optimal for Single-Copy Tomography*. Feb. 2025. arXiv: 2502.18170 [quant-ph] (cit. on p. 34).

[Axl24] Sheldon Axler. *Linear Algebra Done Right*. 4th. Undergraduate Texts in Mathematics. Springer, 2024 (cit. on p. 38).

[Bar+97] Adriano Barenco, André Berthiaume, David Deutsch, et al. "Stabilization of Quantum Computations by Symmetrization". In: *SIAM Journal on Computing* 26.5 (Oct. 1997), pp. 1541–1557 (cit. on p. 81).

[BB87] J. Bertrand and P. Bertrand. "A Tomographic Approach to Wigner's Function". In: *Foundations of Physics* 17.4 (Apr. 1987), pp. 397–405 (cit. on p. 25).

[Buh+01] Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf. "Quantum Fingerprinting". In: *Physical Review Letters* 87.16 (Sept. 2001), p. 167902 (cit. on p. 81).

[CC25] Sitan Chen and Jordan Cotler. *Quantum Learning Theory: Lecture Notes for Harvard Physics 272 / CS 2233*. https://harvard-quantum-learning.github.io/. Fall 2025, Harvard University. 2025 (cit. on pp. 65, 68, 70, 74, 76).

[CDS07] Giulio Chiribella, Giacomo Mauro D'Ariano, and Dirk Schlingemann. "How Continuous Quantum Measurements in Finite Dimensions Are Actually Discrete". In: *Physical Review Letters* 98.19 (May 2007), p. 190403 (cit. on p. 47).

[CGY24] Sitan Chen, Weiyuan Gong, and Qi Ye. *Optimal Tradeoffs for Estimating Pauli Observables*. Apr. 2024. arXiv: 2404.19105 (cit. on pp. 74, 76, 77).

[CGZ24] Sitan Chen, Weiyuan Gong, and Zhihan Zhang. *Adaptivity Can Help Exponentially for Shadow Tomography*. Dec. 2024. arXiv: 2412.19022 [quant-ph] (cit. on p. 77).

[Che+22] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. "Exponential Separations Between Learning With and Without Quantum Memory". In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. Feb. 2022, pp. 574–585 (cit. on pp. 65, 67, 71).

[Die88] Dennis Dieks. “Overlap and distinguishability of quantum states”. In: *Physics Letters A* 126.5 (1988), pp. 303–306 (cit. on p. 10).

[Dir58] Paul A. M. Dirac. *The Principles of Quantum Mechanics*. 4th. First edition published in 1930. Oxford: Clarendon Press, 1958 (cit. on p. 3).

[Eke+02] Artur K. Ekert, Carolina Moura Alves, Daniel K. L. Oi, et al. “Direct Estimations of Linear and Nonlinear Functionals of a Quantum State”. In: *Physical Review Letters* 88.21 (May 2002), p. 217901 (cit. on p. 81).

[Har13] Aram W. Harrow. *The Church of the Symmetric Subspace*. Aug. 2013. arXiv: 1308.6595 (cit. on p. 44).

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. “Predicting Many Properties of a Quantum System from Very Few Measurements”. In: *Nature Physics* 16.10 (Oct. 2020), pp. 1050–1057 (cit. on p. 64).

[HKP21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. “Information-Theoretic Bounds on Quantum Advantage in Machine Learning”. In: *Physical Review Letters* 126.19 (May 2021), p. 190505 (cit. on p. 74).

[Iva87] Igor D. Ivanovic. “How to differentiate between non-orthogonal states”. In: *Physics Letters A* 123.6 (1987), pp. 257–259 (cit. on p. 10).

[Led25] Felix Leditzky. *Representation Theoretic Methods in Quantum Information Theory*. https://felixleditzky.info/teaching/FT25/math595-repth-qit.pdf. Lecture notes for Math 595, University of Illinois Urbana-Champaign. 2025 (cit. on p. 44).

[Leo95] Ulf Leonhardt. “Quantum-State Tomography and Discrete Wigner Function”. In: *Physical Review Letters* 74.21 (May 1995), pp. 4101–4105 (cit. on p. 25).

[LN25] Angus Lowe and Ashwin Nayak. “Lower Bounds for Learning Quantum States with Single-Copy Measurements”. In: *ACM Trans. Comput. Theory* 17.1 (Mar. 2025), 7:1–7:42 (cit. on pp. 34, 62).

[Low21] Angus Lowe. “Learning quantum states without entangled measurements”. MA thesis. University of Waterloo, 2021 (cit. on pp. 34, 50).

[MD16] Ashley Montanaro and Ronald De Wolf. “A Survey of Quantum Property Testing”. In: *Theory of Computing* 1.1 (2016), pp. 1–81 (cit. on p. 78).

[Mel24] Antonio Anna Mele. “Introduction to Haar Measure Tools in Quantum Information: A Beginner’s Tutorial”. In: *Quantum* 8 (May 2024), p. 1340. arXiv: 2307.08956 [quant-ph] (cit. on pp. 35, 44).

[NC00] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000 (cit. on pp. 1, 58).

[Neu55] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Trans. by Robert T. Beyer. Originally published as *Mathematische Grundlagen der Quantenmechanik* in 1932. Princeton, NJ: Princeton University Press, 1955 (cit. on p. 3).

[Nis25] Harumichi Nishimura. “A Survey: SWAP Test and Its Applications to Quantum Complexity Theory”. In: *Algorithmic Foundations for Social Advancement: Recent Progress on Theory and Practice*. Ed. by Shin-ichi Minato, Takeaki Uno, Norihito Yasuda, et al. Singapore: Springer Nature, 2025, pp. 243–261 (cit. on p. 82).

[Per88] Asher Peres. “How to differentiate between non-orthogonal states”. In: *Physics Letters A* 128.1 (1988), p. 19 (cit. on p. 10).

[Sch15] Frederic P. Schuller. *Lectures on Quantum Theory*. https://tales.mbivert.com/Lectures_on_Quantum_Theory_complete.pdf. 2015 (cit. on p. 46).

[Smi+93] D. T. Smithey, M. Beck, M. G. Raymer, and A. Faridani. “Measurement of the Wigner Distribution and the Density Matrix of a Light Mode Using Optical Homodyne Tomography: Application to Squeezed States and the Vacuum”. In: *Physical Review Letters* 70.9 (Mar. 1993), pp. 1244–1247 (cit. on p. 25).

[SSW25] Thilo Scharnhorst, Jack Spilecki, and John Wright. *Optimal Lower Bounds for Quantum State Tomography*. Oct. 2025. arXiv: 2510.07699 [quant-ph] (cit. on p. 62).

[Ste04] J. Michael Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. MAA Problem Books. Cambridge University Press, 2004 (cit. on p. 18).

[Str+22] Roman Stricker, Michael Meth, Lukas Postler, et al. “Experimental Single-Setting Quantum State Tomography”. In: *PRX Quantum* 3.4 (Oct. 2022), p. 040310 (cit. on p. 34).

[Wal18] Michael Walter. *Symmetry and Quantum Information*. https://qi.ruhr-uni-bochum.de/qit18/qit18.pdf. Lecture notes, University of Amsterdam. Last updated May 14, 2023. 2018 (cit. on p. 46).

[Wil17] Mark M. Wilde. *Quantum Information Theory*. 2nd ed. Cambridge, UK: Cambridge University Press, 2017 (cit. on p. 58).

[Wri24] John Wright. *CS 294: Quantum Learning Theory*. Lecture Notes, University of California, Berkeley. 2024 (cit. on pp. 2, 10, 85).

[Yu20] Nengkun Yu. *Sample Efficient Tomography via Pauli Measurements*. Sept. 2020. arXiv: 2009.04610 [quant-ph] (cit. on p. 34).

Index

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.