

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	1
1.1	What is Quantum Learning Theory?	1
2	Quantum State Discrimination	2
2.1	Pure State Discrimination	2
2.1.1	Review: Quantum Measurements	3
2.1.2	Figure of Merit for State Discrimination	4
2.1.3	Trivial Strategies and Limiting Cases	5
2.1.4	Optimal Strategy for Pure State Discrimination	7
2.1.5	Exercises	9
2.2	Mixed State Discrimination	11
2.2.1	Diagonal State Discrimination and Total Variation Distance . .	11
2.2.2	Mixed State Discrimination and Holevo-Helstrom Theorem . .	14
2.2.3	Exercises	18
2.3	Discrimination with Multiple Samples	19
2.3.1	Distinguishing Probability Distributions with Multiple Samples	20
2.3.2	Exercises	24
3	Quantum State Tomography	25
3.1	Measurement Classes	26
3.2	Single-copy, Local Tomography Algorithms	26
3.2.1	Pauli Matrices Crash Course	27
3.2.2	Textbook Pauli Tomography	28
3.2.3	Additional Notes on Textbook Pauli Tomography	32
3.2.4	Project Ideas: Optimized Pauli Tomography	34
3.2.5	Project Idea: Single-setting QST with SIC-POVMs	34
3.3	Representation Theory Crash Course	34
3.3.1	Warm-up: Haar Averages	34
3.3.2	The Church of the Symmetric Subspace	39
4	Solutions to Exercises	40
	Bibliography	44

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 information and computing, please come back and learn some quantum learning theory!

1.1 What is Quantum Learning Theory?

To be written...

Quantum State Discrimination

” *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

¹These 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 $p_i = \text{tr} [\rho E_i]$.

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

$$p_i = \text{tr} [\rho E_i] = \sum_{j=1} \lambda_j \text{tr} [|\psi_j\rangle\langle\psi_j| E_i] = \sum_{j=1} \lambda_j \langle\psi_j| E_i |\psi_j\rangle. \quad (2.1)$$

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

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

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

$$p_i = \text{tr} [\rho \Pi_i]. \quad (2.3)$$

In general, these orthogonal projectors can be expressed as

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

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

$$p_i = \text{tr}[\rho |v_i\rangle\langle v_i|] = \langle v_i | \rho | v_i \rangle. \quad (2.5)$$

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

$$\Pr[\text{Guess 2} | \text{State 1}] = \text{tr}[E_2 |\psi_1\rangle\langle\psi_1|]. \quad (2.6)$$

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

$$\Pr[\text{Guess 1} | \text{State 2}] = \text{tr}[E_1 |\psi_2\rangle\langle\psi_2|]. \quad (2.7)$$

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:

$$P_{\text{worst}} = \max \{ \Pr[\text{Guess 1} | \text{State 2}], \Pr[\text{Guess 2} | \text{State 1}] \} \quad (2.8)$$

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**:

$$P_{\text{minimax}} = \min_{\{E_1, E_2\}} \max \{ \text{tr}[E_1 |\psi_2\rangle\langle\psi_2|], \text{tr}[E_2 |\psi_1\rangle\langle\psi_1|] \}. \quad (2.9)$$

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 states². 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

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

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.

²In 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: $\mathbb{C}^d \rightarrow \mathbb{C}^2$. 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

$$\text{tr}[E\rho] = \text{tr}[E \cdot \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}}], \quad \rho = \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}} \quad (2.11)$$

$$= \text{tr}[\Pi_{\mathcal{S}}E\Pi_{\mathcal{S}} \cdot \rho], \quad \text{cyclicity of trace} \quad (2.12)$$

$$= \text{tr}[E'\rho], \quad (2.13)$$

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 .

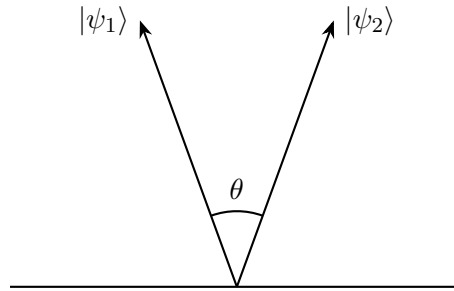


Fig. 2.1: Representation of the two vectors $|\psi_1\rangle$ and $|\psi_2\rangle$. Note that it is without loss of generality to assume $\theta \in [0, \pi/2)$. If this were not the case, one could just replace $|\psi_2\rangle$ with $-|\psi_2\rangle$. States differing by a phase factor are physically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

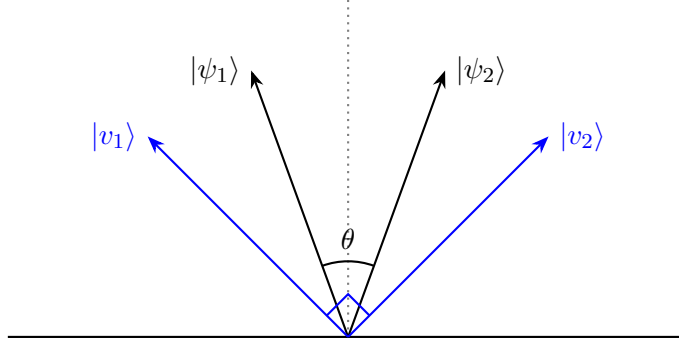


Fig. 2.2: An optimal strategy should not bias one state over the other. Among basis measurements, $\{|v_1\rangle, |v_2\rangle\}$ seems like a promising candidate.

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

$$p_{\text{succ}}(\theta) := \Pr[\text{Guess 1} | \text{State 1}], \quad (2.14)$$

$$= \text{tr} [|v_1\rangle\langle v_1| |\psi_1\rangle\langle\psi_1|], \quad (2.15)$$

$$= |\langle v_1 | \psi_1 \rangle|^2, \quad (2.16)$$

$$= \cos^2 \left(\frac{\pi/2 - \theta}{2} \right), \quad (2.17)$$

$$= \frac{1}{2} + \frac{1}{2} \cos \left(\frac{\pi}{2} - \theta \right), \quad (2.18)$$

$$= \frac{1}{2} + \frac{1}{2} \sin \theta, \quad (2.19)$$

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) **Simulating³ 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 it 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.

³We 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: $p_i = \text{tr}[\rho E_i]$.

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 $\{|\psi_2\rangle, |\psi_2^\perp\rangle\}$ basis.
- (c) **Strategy 3:** Consider the collection of operators $\{E_1, E_2, E_{\text{dk}}\}$ with

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

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

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

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

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

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 $[d] := \{1, 2, \dots, d\}$ which can be written as

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

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?

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

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

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

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

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

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,

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

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

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

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

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.

$$p_{\text{succ}} = \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \sum_{x=1}^2 |p_x - q_x|, \quad (2.32)$$

$$= \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} (|1 - 0| + |0 - 1|), \quad (2.33)$$

$$= 1. \quad (2.34)$$

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 it is given a name!⁴

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

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

⁴See 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.** $d_{\text{TV}}(p, q) \geq 0$, with equality iff $p = q$.
2. **Symmetry.** For any distributions p and q , $d_{\text{TV}}(p, q) = d_{\text{TV}}(q, p)$.
3. **Triangle Inequality.** For any distribution r , we have

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

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 $x = (x_1, \dots, x_d) \in \mathbb{C}^d$, the *vector p -norm* is defined as

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

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

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

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:

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

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:

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

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

$$\|M\|_\infty := \max_i \sigma_i \quad (2.41)$$

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

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

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

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

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

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

$$p_{\text{succ}} = \frac{1}{2} \text{tr}[E_1 \rho] + \frac{1}{2} \text{tr}[E_2 \sigma], \quad \text{equal priors} \quad (2.45)$$

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

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

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

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

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

$$:= P + N, \quad (2.50)$$

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

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

$$= \frac{1}{2} + \frac{1}{2} \text{tr}[E_1 P] + \frac{1}{2} \text{tr}[E_1 N]. \quad (2.52)$$

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

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

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

$$p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2} \text{tr}[P], \quad (2.54)$$

$$= \frac{1}{2} + \frac{1}{2} \sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4} \cdot 0, \quad (2.55)$$

$$= \frac{1}{2} + \frac{1}{2} \sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4} \sum_{i=1}^d \lambda_i, \quad (2.56)$$

$$= \frac{1}{2} + \frac{1}{2} \sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{4} \left(\sum_{i:\lambda_i \geq 0} \lambda_i + \sum_{i:\lambda_i < 0} \lambda_i \right), \quad (2.57)$$

$$= \frac{1}{2} + \frac{1}{2} \left(\frac{1}{2} \sum_{i:\lambda_i \geq 0} \lambda_i - \frac{1}{2} \sum_{i:\lambda_i < 0} \lambda_i \right), \quad (2.58)$$

$$= \frac{1}{2} + \frac{1}{2} \left(\frac{1}{2} \sum_{i:\lambda_i \geq 0} |\lambda_i| + \frac{1}{2} \sum_{i:\lambda_i < 0} |\lambda_i| \right), \quad (2.59)$$

$$= \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \sum_{i=1}^d |\lambda_i|, \quad (2.60)$$

$$= \frac{1}{2} + \frac{1}{2} \|\rho - \sigma\|_1, \quad \text{Def. 2.2.6} \quad (2.61)$$

$$= \frac{1}{2} + \frac{1}{2} d_{\text{tr}}(\rho, \sigma), \quad (2.62)$$

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 $\text{tr}[E_1 N] \leq 0$ and $\text{tr}[E_1 P] \leq \text{tr}[P]$. 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

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

$$= 0, \quad \text{orthonormality of } \{|v_i\rangle\} \quad (2.64)$$

as well as

$$\text{tr}[E_1 P] = \text{tr} \left[\sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i| \sum_{i:\lambda_i \geq 0} \lambda_i |v_i\rangle\langle v_i| \right], \quad (2.65)$$

$$= \text{tr} \left[\sum_{i:\lambda_i \geq 0} \lambda_i |v_i\rangle\langle v_i| \right], \quad (2.66)$$

$$= \sum_{i:\lambda_i \geq 0} \lambda_i, \quad (2.67)$$

$$= \text{tr}[P], \quad (2.68)$$

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 $\rho - \sigma$

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

which yields a maximum probability of success given as

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

□

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

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

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

$$\text{tr}[AB] \leq \|A\|_p \|B\|_q. \quad (2.72)$$

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

$$\mathbf{x} := (x_1, x_2, \dots, x_n) \in \mathbb{R}^n. \quad (2.73)$$

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 \mathbf{x} 's as

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

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

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

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

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

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

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

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, let's 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

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

Recall, also, that

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

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

Given n coin flips, then, we have $x \sim p^{\otimes n}$ 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

$$K = \sum_{i=1}^n X_i, \quad (2.81)$$

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)

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

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

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

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

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

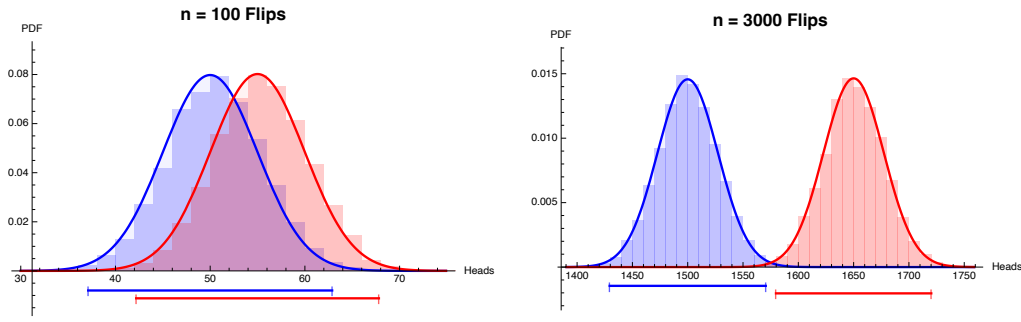


Fig. 2.3: Two histograms showing the distribution of the number of heads given n flips, with the 99% confidence intervals superposed. Placeholder to be replaced with nicer matplotlib plot.

Binomial distribution will approach a Gaussian (normal) distribution⁵ 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

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

or, equivalently,

$$\Pr[|X - \mathbb{E}[X]| \geq c] \leq \frac{\text{Var}[X]}{c^2}. \quad (2.86)$$

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 $K \sim \text{Bin}(n, \alpha)$. We may write

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

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

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

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

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

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

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

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

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

Thus, we need to choose n such

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

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

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

This is an upper bound on the *sample complexity* of distinguishing a fair coin from a slightly biased one. The so-called big-O notation⁶, $O(\cdot)$, 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

$$p' = (1, 0), \quad (2.93)$$

$$q' = (1 - \epsilon, \epsilon). \quad (2.94)$$

⁶If 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

$$d_{\text{TV}}(p, q) = \epsilon, \quad (2.95)$$

$$d_{\text{TV}}(p', q') = \epsilon, \quad (2.96)$$

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 $1 - \delta$.

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

$$\Pr [d_{\text{tr}}(\rho, \hat{\rho}) \leq \epsilon] > 1 - \delta. \quad (3.1)$$

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 $\mathcal{H} = ((\mathbb{C}^d)^{\tilde{\otimes} n})^{\otimes T}$ 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

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

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 satisfy the following properties:

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

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

$$\rho = \frac{1}{2^n} \sum_{i=1}^{d^2} \text{tr}[P_i \rho] P_i, \quad (3.3)$$

where the $P_i \in \{\mathbb{I}, \sigma_x, \sigma_y, \sigma_z\}^{\otimes n}$. The algorithm is simple: measure each coefficient, $\alpha_{P_i} := \text{tr}[P_i \rho]$ 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 measurement¹ $\{\frac{1}{2}(\mathbb{I} + P_i), \frac{1}{2}(\mathbb{I} - P_i)\}$, 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

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

¹This 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

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

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

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

$$= (+1) \cdot \Pr[+1] + (-1) \cdot \Pr[-1], \quad (3.8)$$

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

$$\mathbb{E} [\hat{\alpha}_{P_i}] = \text{tr} [P_i \rho]. \quad (3.10)$$

By linearity 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. $\mathbb{E} [\hat{\rho}] = \rho$). 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:

$$\|A\|_2 \leq \|A\|_1 \leq \sqrt{d} \|A\|_2 \quad (3.11)$$

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 inequality².

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

$$\mathbb{E} [\|\rho - \hat{\rho}\|_2^2]. \quad (3.12)$$

Is this true? Is there an intuitive reason?

²Cauchy-Schwarz... always try Cauchy-Schwarz.

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

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

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

$$= \frac{1}{2^n} \sum_{P \in \mathcal{P}_n} \text{Var} [\hat{\alpha}_P]. \quad (3.15)$$

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 $\mathbb{E} [\|\rho - \hat{\rho}\|_2^2]$ implies an upper bound on $\|\rho - \hat{\rho}\|_1$ with high probability. Observe

$$\mathbb{E} [\|\rho - \hat{\rho}\|_1] \leq \sqrt{\mathbb{E} [\|\rho - \hat{\rho}\|_1^2]}, \quad \text{Var} [X] \geq 0, \quad (3.16)$$

$$\leq \sqrt{d \cdot \mathbb{E} [\|\rho - \hat{\rho}\|_2^2]}, \quad \text{Lemma 3.2.3}, \quad (3.17)$$

$$= \sqrt{d} \sqrt{\mathbb{E} [\|\rho - \hat{\rho}\|_2^2]}. \quad (3.18)$$

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

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

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

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

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

$$\mathbb{E} [\|\rho - \hat{\rho}\|_2^2] \leq \frac{d}{M}. \quad (3.21)$$

Then, the second form of Markov's implies

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

Setting $c = 100$, for example, yields

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

Then, a simple application of Lemma 3.2.3 yields

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

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

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

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

$$\mathbb{E} [\|\hat{\rho} - \rho\|_2^2] = \mathbb{E} \left[\text{tr} \left[(\hat{\rho} - \rho)^\dagger (\hat{\rho} - \rho) \right] \right], \quad (3.26)$$

$$= \mathbb{E} \left[\frac{1}{d^2} \sum_{i=1}^{d^2} \sum_{j=1}^{d^2} (\hat{\alpha}_{P_i} - \text{tr}[P_i \rho]) (\hat{\alpha}_{P_j} - \text{tr}[P_j \rho]) \underbrace{\text{tr}[P_i P_j]}_{d\delta_{ij}} \right], \quad (3.27)$$

$$= \mathbb{E} \left[\frac{1}{d} \sum_{i=1}^{d^2} (\hat{\alpha}_{P_i} - \text{tr}[P_i \rho])^2 \right], \quad (3.28)$$

$$= \frac{1}{d} \sum_{i=1}^{d^2} \text{Var} [\hat{\alpha}_{P_i}], \quad (3.29)$$

$$= \frac{1}{d} \sum_{i=1}^{d^2} \text{Var} \left[\frac{1}{M} \sum_{m=1}^M x_i^{(m)} \right], \quad (3.30)$$

$$= \frac{1}{dM^2} \sum_{i=1}^{d^2} \sum_{m=1}^M \text{Var} [x_i^{(m)}], \quad (3.31)$$

$$= \frac{1}{dM^2} \sum_{i=1}^{d^2} \sum_{m=1}^M \left(\mathbb{E} [(x_i^{(m)})^2] - \mathbb{E} [x_i^{(m)}]^2 \right), \quad (3.32)$$

$$= \frac{1}{dM^2} \sum_{i=1}^{d^2} \sum_{m=1}^M \left(1 - \mathbb{E} [x_i^{(m)}]^2 \right), \quad (3.33)$$

$$\leq \frac{1}{dM^2} \sum_{i=1}^{d^2} \sum_{m=1}^M 1, \quad (3.34)$$

$$\implies \mathbb{E} [\|\hat{\rho} - \rho\|_2^2] \leq \frac{d}{M}. \quad (3.35)$$

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

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

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 $b = \prod_{i=1}^n b_i$.

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

$$\Pr_{\text{global}}[b] = \Pr_{\text{local}}[b]. \quad (3.37)$$

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

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

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

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

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

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

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

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

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

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

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

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

where we have defined

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

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

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

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

Plugging this back into our expression, we have

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

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

$$= \frac{1 + b \cdot \text{tr}[\rho P]}{2}, \quad (3.50)$$

$$= \Pr_{\text{global}}[b], \quad (3.51)$$

as desired. \square

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: Optimized Pauli Tomography

3.2.5 Project Idea: Single-setting QST with SIC-POVMs

3.3 Representation Theory Crash Course

3.3.1 Warm-up: Haar Averages

Let us denote the unitary group $\mathcal{U}_d := \{U \in \mathcal{L}(\mathbb{C}^d) : U^\dagger U = \mathbb{I}_d\}$. 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

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

We note that for any (measurable) $S \subseteq \mathcal{U}_d$, it follows that $\int_S 1 dU \geq 0$ and $\int_{\mathcal{U}_d} 1 dU = 1$, 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

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

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.

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

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

$$\text{tr}[A] = \text{tr} \left[\mathbb{E}_{U \sim \mu_H} [UOU^\dagger] \right], \quad (3.55)$$

$$= \text{tr} \left[\int_{U(d)} UOU^\dagger d\mu(U) \right], \quad (3.56)$$

$$= \int_{U(d)} \text{tr} [UOU^\dagger] d\mu(U), \quad (3.57)$$

$$= \int_{U(d)} \text{tr} [O] d\mu(U), \quad \text{cyclicity} \quad (3.58)$$

$$= \text{tr} [O], \quad (3.59)$$

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,

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

which we may write suggestively as

$$\mathbb{E}_{U \sim \mu_H} [U|0\rangle\langle 0|U^\dagger] = \frac{\text{tr}[|0\rangle\langle 0|]}{2} \cdot \mathbb{I}. \quad (3.61)$$

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, $\langle \psi | U Z U^\dagger | \psi \rangle$ 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.

$$\mathbb{E}_{U \sim \mu_H} [U Z U^\dagger] = \mathbb{E}_{U \sim \mu_H} [U|0\rangle\langle 0|U^\dagger] - \mathbb{E}_{U \sim \mu_H} [U|1\rangle\langle 1|U^\dagger] = 0, \quad (3.62)$$

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

$$\mathbb{E}_{U \sim \mu_H} [U \mathbb{I} U^\dagger] = \mathbb{E}_{U \sim \mu_H} [U U^\dagger] = \mathbb{E}_{U \sim \mu_H} [\mathbb{I}] = \mathbb{I}, \quad (3.63)$$

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

Proposition 3.3.5 (Haar Average of an Operator).

$$\mathbb{E}_{U \sim \mu_H} [U O U^\dagger] = \frac{\text{tr}[O]}{d} \mathbb{I}. \quad (3.64)$$

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

$$V A V^\dagger = V \left(\mathbb{E}_{U \sim \mu_H} [U O U^\dagger] \right) V^\dagger, \quad (3.65)$$

$$= V \left(\int_{\mathcal{U}_d} U O U^\dagger d\mu(U) \right) V^\dagger, \quad (3.66)$$

$$= \left(\int_{\mathcal{U}_d} V U O U^\dagger V^\dagger d\mu(U) \right), \quad (3.67)$$

$$= \left(\int_{\mathcal{U}_d} V U O (V U)^\dagger d\mu(U) \right), \quad (3.68)$$

$$= \left(\int_{\mathcal{U}_d} U O U d\mu(U) \right), \quad \text{left-invariance of Haar measure} \quad (3.69)$$

$$= A, \quad (3.70)$$

which implies $V A = A V$ 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

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

Hermitian operators have real eigenvalues. Further, we can restrict our attention to operators satisfying $\|H\|_\infty \leq 1$. 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

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

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

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

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

$$\text{tr} \left[\mathbb{E}_{U \sim \mu_H} [U O U^\dagger] \right] = \text{tr} [\lambda \mathbb{I}], \quad (3.74)$$

$$\mathbb{E}_{U \sim \mu_H} [\text{tr} [U O U^\dagger]] = \lambda \cdot d, \quad (3.75)$$

$$\mathbb{E}_{U \sim \mu_H} [\text{tr} [O]] = \lambda \cdot d, \quad \text{cyclicity of trace} \quad (3.76)$$

$$\implies \lambda = \frac{\text{tr} [O]}{d}, \quad (3.77)$$

as desired. \square

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.

³This 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.

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

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

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

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 $\mathcal{M}_{\mu_H}^{(1)}(O)$ 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 $\mathcal{M}_{\mu_H}^{(1)}(O)$ is one dimensional, i.e. spanned by \mathbb{I} .

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

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

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 $\mathcal{M}_{\mu_H}^{(2)}(O)$. 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

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 $\{|v_j\rangle\}$ such that

$$A = \sum_{i=1}^d p_i |u_i\rangle \langle u_i| \quad \text{and} \quad B = \sum_{j=1}^d q_j |v_j\rangle \langle v_j|. \quad (4.1)$$

This allows us to write

$$\text{tr}[AB] = \text{tr} \left[\sum_i p_i |u_i\rangle \langle u_i| \sum_j q_j |v_j\rangle \langle v_j| \right], \quad (4.2)$$

$$= \sum_{ij} p_i q_j \text{tr} [|u_i\rangle \langle u_i| v_j\rangle \langle v_j|], \quad (4.3)$$

$$= \sum_{ij} p_i q_j |\langle u_i | v_j \rangle|^2, \quad (4.4)$$

$$= \sum_{ij} p_i q_j \left(|\langle u_i | v_j \rangle|^2 \right)^1, \quad (4.5)$$

$$= \sum_{ij} p_i q_j \left(|\langle u_i | v_j \rangle|^2 \right)^{\frac{1}{p} + \frac{1}{q}}, \quad (4.6)$$

$$= \sum_{ij} \left(p_i |\langle u_i | v_j \rangle|^{\frac{2}{p}} \right) \left(q_j |\langle u_i | v_j \rangle|^{\frac{2}{q}} \right), \quad (4.7)$$

$$\leq \left(\sum_{ij} p_i^p |\langle u_i | v_j \rangle|^2 \right)^{\frac{1}{p}} \left(\sum_{ij} q_j^q |\langle u_i | v_j \rangle|^2 \right)^{\frac{1}{q}}, \quad \text{Hölder's Inequality} \quad (4.8)$$

$$= \left(\sum_i p_i^p \sum_j |\langle u_i | v_j \rangle|^2 \right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i |\langle u_i | v_j \rangle|^2 \right)^{\frac{1}{q}}, \quad (4.9)$$

$$= \left(\sum_i p_i^p \sum_j \langle u_i | v_j \rangle \langle v_j | u_i \rangle \right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i \langle v_j | u_i \rangle \langle u_i | v_j \rangle \right)^{\frac{1}{q}}, \quad (4.10)$$

$$= \left(\sum_i p_i^p \right)^{\frac{1}{p}} \left(\sum_j q_j^q \right)^{\frac{1}{q}}, \quad (4.11)$$

$$= \|A\|_p \|B\|_q, \quad (4.12)$$

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. \square

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

$$p_{\text{succ}} = \frac{1}{2} \text{tr} [E_1 \rho_1] + \frac{1}{2} \text{tr} [\rho_2 E_2], \quad (4.13)$$

$$= \frac{1}{2} \text{tr} [E_1 \rho_1 + \rho_2 E_2], \quad (4.14)$$

$$= \frac{1}{4} \text{tr} [E_1 \rho_1 + \rho_2 E_2] + \frac{1}{4} \text{tr} [E_1 \rho_1 + \rho_2 E_2], \quad (4.15)$$

$$= \frac{1}{4} \text{tr} [(E_1 + E_2)(\rho_1 + \rho_2)] + \frac{1}{4} \text{tr} [(E_1 - E_2)(\rho_1 - \rho_2)], \quad (4.16)$$

$$= \frac{1}{2} + \frac{1}{4} \text{tr} [(E_1 - E_2)(\rho_1 - \rho_2)], \quad (4.17)$$

$$=: \frac{1}{2} + \frac{1}{2} T, \quad (4.18)$$

$$(4.19)$$

where we have used that $E_1 + E_2 = \mathbb{I}$ and $\text{tr} [\rho_i] = 1$. Now, we can apply Holder's inequality

$$T = \frac{1}{2} \text{tr} [(E_1 - E_2)(\rho_1 - \rho_2)], \quad (4.20)$$

$$\leq \frac{1}{2} \|E_1 - E_2\|_{\infty} \|\rho_1 - \rho_2\|_1, \quad (4.21)$$

$$\leq \frac{1}{2} \|\rho_1 - \rho_2\|_1, \quad (4.22)$$

$$= d_{\text{tr}}(\rho_1, \rho_2) \quad (4.23)$$

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

$$p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2} d_{\text{tr}}(\rho_1, \rho_2), \quad (4.24)$$

as desired. We won't show the optimal measurement strategy, because it is the same as above. \square

Bibliography

- [Axl24] Sheldon Axler. *Linear Algebra Done Right*. 4th. Undergraduate Texts in Mathematics. Springer, 2024 (cit. on p. 38).
- [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).
- [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).
- [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).
- [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 p. 34).
- [Low21] Angus Lowe. “Learning quantum states without entangled measurements”. MA thesis. University of Waterloo, 2021 (cit. on p. 34).
- [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 p. 34).
- [NC00] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000 (cit. on p. 1).
- [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).
- [Per88] Asher Peres. “How to differentiate between non-orthogonal states”. In: *Physics Letters A* 128.1 (1988), p. 19 (cit. on p. 10).
- [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).

- [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).
- [Wri24] John Wright. *CS 294: Quantum Learning Theory*. [Lecture Notes](#), University of California, Berkeley. 2024 (cit. on pp. 2, 10, 40).

Index

k -th moment operator, [39](#)

Average-case error, [11](#)

Basis measurement, [4](#)

Bernoulli Random Variable, [21](#)

Chebyshev's Inequality, [22](#)

Commutant, [39](#)

Concentration Inequality, [22](#)

Hölder's Inequality for Matrices, [18](#)

Haar Measure, [34](#)

Haar random state, [36](#)

Holevo-Helstrom Theorem, [15](#)

Markov's Inequality, [30](#)

Pauli Matrices, [27](#)

Positive Operator-valued Measure, [3](#)

Projection-valued Measure, [3](#)

Quantum State Discrimination, [2](#)

Schatten p -norm, [15](#)

Symmetric Subspace, [39](#)

Total Variation Distance, [13](#)

Trace Distance, [15](#)

Unambiguous State Discrimination, [9](#)

Vector p -norm, [14](#)

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/>.