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

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

Because $\rho \geq 0$, we know $\lambda_j \geq 0$. Thus, for p_i to be non-negative, we require $\left\langle\psi_j\right| E_i \left|\psi_j\right\rangle \geq 0$ 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

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

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 $\mathcal{S} = \mathrm{span}\{|\psi_1\rangle, |\psi_2\rangle\}$. 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

$$\mathrm{tr}\left[E\rho\right] = \mathrm{tr}\left[E \cdot \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}}\right], \quad \rho = \Pi_{\mathcal{S}}\rho\Pi_{\mathcal{S}} \tag{2.11}$$

$$= \mathrm{tr}\left[\Pi_{\mathcal{S}}E\Pi_{\mathcal{S}} \cdot \rho\right], \quad \text{cyclicity of trace} \tag{2.12}$$

$$= \mathrm{tr}\left[E'\rho\right], \tag{2.13}$$

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 $\mathbb{C}^2$.

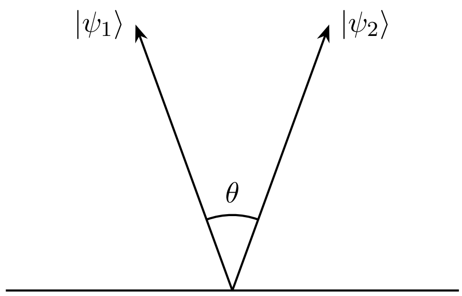

Fig. 2.1: Representation of the two vectors $|\psi_1\rangle$ and $|\psi_2\rangle$. Note that it is without loss of generality to assume $\theta \in [0, \pi/2)$. If this were not the case, one could just replace $|\psi_2\rangle$ with $-|\psi_2\rangle$. States differing by a phase factor are physically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

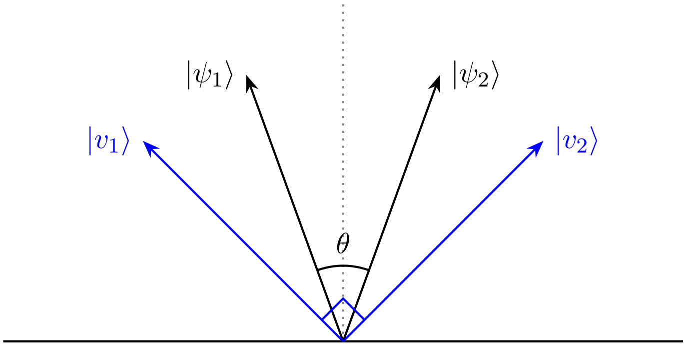

Fig. 2.2: An optimal strategy should not bias one state over the other. Among basis measurements, $\{|v_1\rangle, |v_2\rangle\}$ seems like a promising candidate.

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

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

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

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

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

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

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

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

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

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

as well as

$$\operatorname{tr}[E_1 P] = \operatorname{tr}\left[\sum_{i:\lambda_i \geq 0} |v_i\rangle\langle v_i| \sum_{i:\lambda_i \geq 0} \lambda_i |v_j\rangle\langle v_j|\right], \tag{2.65}$$

$$= \operatorname{tr}\left[\sum_{i:\lambda_i \geq 0} \lambda_i |v_j\rangle\langle v_j|\right], \tag{2.66}$$

$$= \sum_{i:\lambda_i \geq 0} \lambda_i, \tag{2.67}$$

$$= \operatorname{tr}[P], \tag{2.68}$$

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 $\rho - \sigma$

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

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

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

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 Chebyshev and a Sample Complexity Upper Bound

> **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 $X \sim \mathrm{Bern}(\alpha)$, 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

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

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

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 $X_i \sim \mathrm{Bern}(\alpha)\ \forall\, i \in [n]$. 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)

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

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

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

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

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

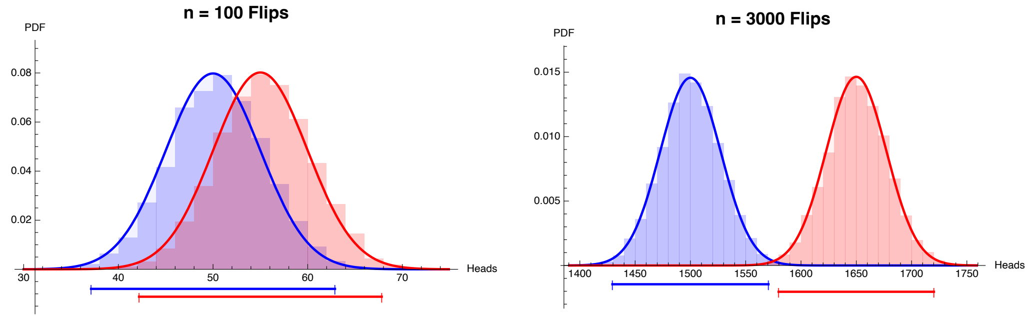

Fig. 2.3: Two histograms showing the distribution of the number of heads given n flips, with the 99% confidence intervals superposed. Placeholder to be replaced with nicer matplotlib plot.

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

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

or, equivalently,

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

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

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

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, I highly recommend you watch this video before proceeding!

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, but intuitively, the idea is that if you have p', you are guaranteed to get all heads, but if you have q', the probability of getting all heads after n flips is

$$\Pr[\text{all heads}] = (1-\epsilon)^n \approx 1 - n\epsilon, \tag{2.95}$$

which is approximately zero when $n = O(1/\epsilon)$. This highlights a shortcoming of the total variation distance. Notice that

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

$$d_{\mathrm{TV}}(p', q') = \epsilon, \tag{2.97}$$

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 $\delta \in (0, 1)$. Use Chebyshev's inequality to find how many samples suffice to distinguish between $p'^{\otimes n}$ and $q'^{\otimes n}$ with probability at least $1 - \delta$.

3 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

$$A = \sum_{i=1}^{d} p_i |u_i\rangle \langle u_i| \quad \text{and} \quad B = \sum_{j=1}^{d} q_j |v_j\rangle \langle v_j| . \tag{3.1}$$

This allows us to write

$$\operatorname{tr}[AB] = \operatorname{tr}\left[\sum_i p_i |u_i\rangle\langle u_i| \sum_j q_j |v_j\rangle\langle v_j|\right], \tag{3.2}$$

$$= \sum_{ij} p_i q_j \operatorname{tr}[|u_i\rangle\langle u_i|v_j\rangle\langle v_j|], \tag{3.3}$$

$$= \sum_{ij} p_i q_j |\langle u_i|v_j\rangle|^2, \tag{3.4}$$

$$= \sum_{ij} p_i q_j \left(|\langle u_i|v_j\rangle|^2\right)^1, \tag{3.5}$$

$$= \sum_{ij} p_i q_j \left(|\langle u_i|v_j\rangle|^2\right)^{\frac{1}{p}+\frac{1}{q}}, \tag{3.6}$$

$$= \sum_{ij} \left(p_i |\langle u_i|v_j\rangle|^{\frac{2}{p}}\right)\left(q_j |\langle u_i|v_j\rangle|^{\frac{2}{q}}\right), \tag{3.7}$$

$$\leq \left(\sum_{ij} p_i^p |\langle u_i|v_j\rangle|^2\right)^{\frac{1}{p}} \left(\sum_{ij} q_j^q |\langle u_i|v_j\rangle|^2\right)^{\frac{1}{q}}, \quad \text{Hölder's Inequality} \tag{3.8}$$

$$= \left(\sum_i p_i^p \sum_j |\langle u_i|v_j\rangle|^2\right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i |\langle u_i|v_j\rangle|^2\right)^{\frac{1}{q}}, \tag{3.9}$$

$$= \left(\sum_i p_i^p \sum_j \langle u_i|v_j\rangle\langle v_j|u_i\rangle\right)^{\frac{1}{p}} \left(\sum_j q_j^q \sum_i \langle v_j|u_i\rangle\langle u_i|v_j\rangle\right)^{\frac{1}{q}}, \tag{3.10}$$

$$= \left(\sum_i p_i^p\right)^{\frac{1}{p}} \left(\sum_j q_j^q\right)^{\frac{1}{q}}, \tag{3.11}$$

$$= \|A\|_p \|B\|_q, \tag{3.12}$$

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. □

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

$$
\begin{aligned}
p_{\text{succ}} &= \frac{1}{2}\operatorname{tr}\left[E_1\rho_1\right] + \frac{1}{2}\operatorname{tr}\left[\rho_2 E_2\right], && (3.13)\\
&= \frac{1}{2}\operatorname{tr}\left[E_1\rho_1 + \rho_2 E_2\right], && (3.14)\\
&= \frac{1}{4}\operatorname{tr}\left[E_1\rho_1 + \rho_2 E_2\right] + \frac{1}{4}\operatorname{tr}\left[E_1\rho_1 + \rho_2 E_2\right], && (3.15)\\
&= \frac{1}{4}\operatorname{tr}\left[(E_1+E_2)(\rho_1+\rho_2)\right] + \frac{1}{4}\operatorname{tr}\left[(E_1-E_2)(\rho_1-\rho_2)\right], && (3.16)\\
&= \frac{1}{2} + \frac{1}{4}\operatorname{tr}\left[(E_1-E_2)(\rho_1-\rho_2)\right], && (3.17)\\
&=: \frac{1}{2} + \frac{1}{2}T, && (3.18)\\
& && (3.19)
\end{aligned}
$$

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

$$
\begin{aligned}
T &= \frac{1}{2}\operatorname{tr}\left[(E_1-E_2)(\rho_1-\rho_2)\right], && (3.20)\\
&\leq \frac{1}{2}\|E_1 - E_2\|_\infty \|\rho_1-\rho_2\|_1, && (3.21)\\
&\leq \frac{1}{2}\|\rho_1-\rho_2\|_1, && (3.22)\\
&= d_{\text{tr}}(\rho_1,\rho_2) && (3.23)
\end{aligned}
$$

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

$$
p_{\text{succ}} \leq \frac{1}{2} + \frac{1}{2}d_{\text{tr}}(\rho_1,\rho_2), \qquad (3.24)
$$

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

Bibliography

[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).

[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).

[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, 25).

Index

Colophon

This thesis was typeset with $\LaTeX\,2_\varepsilon$. 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/`.