Preface

These are lecture notes for a course on quantum learning theory (Physics 172 / CS 2233: Quantum Learning Theory). We begin our journey by motivating our studies in two different ways, one more pragmatic and one more philosophical.

Machine learning, and in particular its recent manifestation in deep learning in the last two decades, has been transformative for computer science and information technology. The promise, perils, and possibility of generative artificial intelligence have seeped from Silicon Valley to the public discourse, and the ultimate contours of its potential are the subject of intense speculation. Granted all of the recent developments in contemporary machine learning, many of the core ideas derive from *statistical learning theory*, which had its heyday in the 1990's and early 2000's. This is a rigorous mathematical subject which conceives of learning in a probabilistic and often Bayesian manner, drawing on probability theory and empirical process theory, while utilizing information-theoretic concepts from Shannon's foundational work. Since contemporary machine learning is mostly an *empirical* subject pertaining to extraordinarily sophisticated statistical models which defy comprehensive characterization, the particularities of the theorems developed in statistical learning theory are not often used; however, the intuitions these rigorous results provide are essential for designing new neural network architectures, loss functions, training algorithms, and datasets. As such, the afterlife of statistical learning theory is that its quantitative knowledge in mathematically simple settings has been lifted to qualitative but indispensable wisdom about highly complex systems.

Then one motivation for our studies is to develop a quantum version of statistical learning theory (or more succinctly, quantum learning theory), suitable for future application by quantum computers. Our studies will focus on quantum learning for *quantum data* as opposed to classical data, for reasons that will be explained. (Indeed, the latter setting is very interesting but has a somewhat different character.) The subject will necessarily be mathematically rigorous to cement our understanding of quantum data and quantum learning algorithms, as well as to develop robust methods with provable performance guarantees suitable for scientific applications. We emphasize that at this moment in time, quantum learning theory is not chiefly an empirical subject such as contemporary machine learning; this underscores the necessity of mathematical rigor and the importance of the foundational development of basic quantum learning algorithms and methods that future theoretical or empirical inquiries may build on. We will focus on developments in quantum learning theory mostly from 2019 onward, which saw the development of fruitful foundations and applications of the subject.

There is also a second, more philosophical motivation for our study of quantum learning theory. *Epistemology* is the philosophical study of what we can know about the world, and how we come to know it. One of the earlier treatments of the subject

goes back over 2000 years to Plato, although among scientists Descartes' maxim "I think therefore I am" (*cogito ergo sum*) may be more familiar. Specifically, Descartes was concerned with what he could know with certainty about the world, and upon wrestling with various uncertainties he concludes that he knows at the very least that he himself exists, since for him to even render the thought requires his own existence.

A persistent thread in epistemology since the beginning is that there may be aspects of our reality that we can never come to know. A particularly incisive analysis along these lines was developed by Immanuel Kant in the late 18th century, in which he detailed how the physicality of our corporeal beings and the constitution of our minds place a priori fundamental limitations on what we can know about the world, leaving certain truths necessarily out of our reach. While this premise is widely accepted by philosophers, it is often frowned upon by scientists; after all, we are children of the Enlightenment for which scientific knowledge is infinitely extensible and far-reaching. If you feel such an urge to frown on epistemology, consider this more modern example: we live in a universe which is expanding and accelerating. Eventually, the expansion will be so fast that light from the early universe will become so redshifted as to be undetectable. As such, if there is life that develops somewhere in the universe at such a time, they will never be able to empirically determine that there was a Big Bang. Thus a truth about the universe is, to them, out of reach.

In the early 20th century, David Hilbert famously declared that the *mathematical* world was fundamentally knowable, and that every precise mathematical statement was either true or false. This epistemic totalism was shockingly undermined by Kurt Gödel in 1931, when he showed that there must always exist mathematical statements which can neither be proved true nor false. This death blow to Hilbert's (and Bertrand Russell's) conception of mathematical knowledge was concretized by Alan Turing in his foundational work on computer science, the pragmatic heir to mathematical logic. Turing famously showed in 1936 that there is no algorithm (which is guaranteed to terminate in finite time) that can conclusively decide if any given algorithm will halt or not. Thus Turing's theory of undecidability cleaves out facts about the world which are fundamentally unknowable to us, furnishing totally precise examples of epistemic roadblocks.

Decades later starting in the early 1970's, the subject of *computational complexity* began to emerge. Instead of being concerned with whether the solution to a computational problem was knowable or unknowable, the subject focused on the *difficulty* of computational problems. For example, one can show that sorting a list of n items (on a classical computer) requires *at most* $\sim n \log n$ computational steps, but also *at least* $\sim n \log n$ computational steps, thus pinpointing the absolute difficulty of the problem. Some computational problems have polynomial difficulty whereas others have exponential difficulty, and are stratified according to *computational complexity classes*. In this way, computational complexity theory comprises a quantitative form of epistemology, circumscribing how difficult it is to obtain computational knowledge.

Having set the scene, we turn to a deep question: how do we come to learn about the world through scientific inquiry? A key facet is that we interrogate the natural world through experiment, and algorithmically process our collected data to reveal hitherto unknown properties of nature. More formally, we can conceptualize

a system in nature – such as a superconductor, a vat of chemicals, a biological organism, etc. – as a source of *data* which is not fully characterized (or else we would not need to run the experiment); then our experiment comprises a series of interactions with the world to sample data, subsequent computational processing of the data, and possibly additional interactions with the world predicated on the processing of previous data. The ultimate outcome is that we learn a property of the world, such as the charge of the electron, the symmetry of a crystal, etc. In this manner, we see that scientific experiments can be beautifully and precisely abstracted into the framework of learning theory. Therefore, a quantitative study of learning theory can reveal what we can fundamentally come to know about the world, and how difficult it is to do so.

Since the laws of nature are quantum-mechanical, any theory of learning the natural world must take quantum mechanics into account. In particular, the natural systems we seek to understand may be quantum-mechanical; the data we extract can be quantum-mechanical; and our means of processing that data can be quantum-mechanical. Thus we necessitate a quantum theory of learning. Such a theory reveals that there are facets of the natural world which are inaccessible to us unless we can harness quantum computers to couple to natural systems and perform quantum information processing. More bluntly, quantum learning gives us access to properties of the natural world which are otherwise unknowable by classical means. And yet the same theory shows us which properties of the natural world are forever out of reach, even with the aid of vast quantum computational power.

Quantum learning theory circumscribes what is knowable and unknowable about the natural world, providing a quantitative epistemology of the grasp of scientific inquiry. With so much at stake, let us begin.

CHAPTER 1

A Sneak Peek: Learning a Rotation Matrix

Before we dive into the formalism of quantum learning, let us begin with a simple motivating example. You will not need to know any quantum mechanics to understand the setup, but it mirrors, in a stripped–down way, how many real experiments try to learn an unknown quantum process that one can interact with in the lab.

1. Basic Setup

Suppose there is an unknown two-dimensional rotation matrix

$$U = R(\theta) \triangleq \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta)\,, \end{pmatrix}$$

and for simplicity assume that $0 \leq \theta < \pi/2$. Your goal is to figure out what θ is, to within small error.

You can "perform an experiment" on it via the following model. Starting from the first standard basis vector $v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, you can decide in advance upon any collection of "controls" specified by rotation matrices $O_0 = R(\theta_0), \ldots, O_m = R(\theta_m)$, and apply (i.e., left-multiply by) the transformation

$$O_m U O_{m-1} U O_{m-2} \cdots U O_0\,.$$

This results in some new unit vector $w = \begin{bmatrix} x \\ y \end{bmatrix}$.

Our figure of merit will be **statistical efficiency**, namely we want to learn θ to within some acceptable level of error using as few experiments and "queries" to U (e.g., the above experiment makes m queries to U) as possible.

If we could see the entries of w, then it's not hard to learn U. In fact we don't even need the full flexibility of picking $O_1, \ldots, O_m$: we can simply take $m = 0$ and use no controls whatsoever, so that the above transformation is given by U itself and $w = Uv$. In this case, if $w = (x, y)$, we can simply read off the angle of rotation defining U from $\theta = \arccos(x)$.

There is a crucial catch however: in physical experiments, we never get to see the literal vector w resulting from an experiment. Without getting into the quantum details yet, the reason is that w is a *superposition* between two different states, namely the first standard basis vector and the second standard basis vector. What we can do is **measure** w, at which point we observe one state or the other, but in a probabilistic fashion.

Definition 1 (Born rule – baby version)**.** *Given unit vector $w = (x, y)$, if we* measure *it, we get as output either $\begin{bmatrix}1\\0\end{bmatrix}$ with probability x^2, or $\begin{bmatrix}0\\1\end{bmatrix}$ with probability y^2. Note that as w is a unit vector, this is a valid probability distribution.*

One might thus envision a natural workaround to not having access to the exact entries (x, y) of w. Measuring effectively gives us access to biased coin flips: with probability x^2 we see heads and with probability y^2 we see tails. By repeatedly performing the experiment that results in w and measuring w each time, we can estimate x^2 and y^2 simply by computing the fraction of heads and tails we observe. How many repetitions do we need?

This can be computed with the **Chernoff bound**:

Fact 2 (Chernoff bound)**.** *Let $X_1, \ldots, X_N$ be independent Bernoulli random variables with expectation p. Then for any $t > 0$,*

$$\Pr\Big[\Big|\frac{1}{N}\sum_i X_i - p\Big| > t\Big] \le 2\exp\Big(-\frac{Nt^2}{2p(1-p)}\Big)$$

In our setting, $p = x^2$, and $\hat{X} = \frac{1}{N}\sum_i X_i$, so if we apply the above with $t = \epsilon x\sqrt{1-x^2}$ and $N = 2\log(2/\delta)/\epsilon^2$, the right-hand side of the above bound is δ and we conclude that with

$$O(\log(2/\delta)/\epsilon^2) \tag{1}$$

coin tosses, we can produce an estimate $0 \le \hat{X} \le 1$ such that $|x^2 - \hat{X}| \le \epsilon x\sqrt{1-x^2}$ with probability at least $1-\delta$. We can then output $\arccos(\sqrt{\hat{X}})$ and argue that this is $O(\epsilon)$-close to θ with some elementary calculus (the reader can safely skip this on a first reading without losing any of the core intuition):

Proposition 3. *Let $0 \le \epsilon \le 1/2$. Suppose $0 \le X, X' \le 1$ satisfy $|X - X'| \le \epsilon\sqrt{X(1-X)}$. Then*

$$|\arccos(\sqrt{X}) - \arccos(\sqrt{X'})| \le \epsilon\,.$$

Proof. As $\arccos(\sqrt{X})$ and $\sqrt{X(1-X)}$ are symmetric about $X = 1/2$, we may assume without loss of generality that $|X| \le 1/2$.

If $|X| \ge \frac{\epsilon^2}{\epsilon^2+4}$, then $X/2 \le X' \le X + \epsilon/2$. If we define $f(z) \triangleq \arccos(\sqrt{z})$, then $f'(z) = -\frac{1}{2\sqrt{z(1-z)}}$, so $|f'(z)| \le 2|f'(X)|$ for all z between X and X'. By integrating, we conclude that

$$|\arccos(\sqrt{X}) - \arccos(\sqrt{X'})| \le \frac{1}{\sqrt{X(1-X)}} \cdot \epsilon\sqrt{X(1-X)} \le \epsilon$$

as desired.

If $|X| < \frac{\epsilon^2}{\epsilon^2+4} \le \frac{\epsilon^2}{4}$, then $\arccos(\sqrt{X}) \ge \pi/2 - \epsilon$. If $X' \le X$, then $\arccos(\sqrt{X}) \le \arccos(\sqrt{X'}) \le \pi/2$, so $|\arccos(\sqrt{X}) - \arccos(\sqrt{X'})| \le \epsilon$. If $X' \ge X$, then $|f'(z)| \le 2|f'(X)|$ for all z between X' and X, so the claimed bound follows again by integrating. □

The $1/\epsilon^2$ scaling in Eq. (1) for the number of coin tosses is called the **standard quantum limit** – often it is formulated in the reverse direction, namely using N experiments (sometimes called "shots"), one can estimate the unknown parameter θ to error $\sim 1/\sqrt{N}$.

2. Beating the Standard Quantum Limit

Of course, we are not yet done. While it is an unavoidable fact of life that in the classical world, estimating the bias of a coin to error ϵ requires $\sim 1/\epsilon^2$ coin tosses in general, in the quantum world we are not limited to simply reducing learning rotations to learning the bias of a random coin. Indeed, the approach described above is exceedingly naive: we set $m = 0$ and didn't use any controls O_i whatsoever.

It turns out that by being clever about the choice of experiments, we can do much better, in fact with only $O(\log 1/\epsilon)$ experiments and $O(1/\epsilon)$ queries to U in total across all experiments. The $O(1/\epsilon)$ scaling is called the **Heisenberg limit**: this turns out to be a fundamental barrier that no experimental protocol, no matter how clever, can beat.

The key idea is to "bootstrap." Instead of estimating θ to high precision right off the bat, we are going to gradually refine our estimate. As a thought experiment, imagine we start by getting a relatively crude approximation to θ by running the protocol in the previous section for target precision which is just a small constant, say, $\epsilon_{\mathsf{crude}} = 1/4$. We can accomplish this with probability $1 - \delta$ using only $O(\log(1/\delta))$ experiments and queries to U, with no dependence yet on the final target precision ϵ.

Given this estimate, if we further subtract $\epsilon_{\mathsf{crude}}$ from it, we get an angle $\theta^{(1)}$ which is an underestimate of θ by a margin of at most $2\epsilon_{\mathsf{crude}} \leq 1/2$. To estimate θ, it now suffices to estimate the *residual angle* $\theta - \theta^{(1)}$. So in all subsequent experiments, instead of querying $U = R(\theta)$, we can query

$$U^{(1)} \triangleq UR(\theta^{(1)})^\dagger = R(\theta - \theta^{(1)}) .$$

Here is our main claim:

Lemma 4. *Suppose $\theta - 2^{-k} \leq \theta^{(k)} \leq \theta$ and let $U^{(k)} \triangleq R(\theta - \theta^{(k)})$. Let $\delta_k > 0$. Consider the following protocol:*

- *Repeat the following experiment $C \log 1/\delta_k$ times:*
  - *Apply $U^{(k)}$ a total of 2^k times starting from the first standard basis vector v.*
  - *Measure the resulting vector and record the observation (heads or tails)*
- *Let $\hat{X}$ denote the fraction of heads seen across these experiments.*
- *Define*

$$\theta^{(k+1)} = \theta^{(k)} + \arccos(\sqrt{\hat{X}})/2^k - 1/2^{k+2} .$$

For C a sufficiently large absolute constant, we have $\theta - 1/2^{k+1} \leq \theta^{(k+1)} \leq \theta$ with probability at least $1 - \delta_k$.

Proof sketch. Note that the rotation given by applying $U^{(k)}$ a total of 2^k times is $R(2^k(\theta - \theta^{(k)}))$. By taking C sufficiently large, the argument in the previous section implies that $|\arccos(\sqrt{\hat{X}}) - 2^k(\theta - \theta^{(k)})| \leq 1/4$. Dividing by 2^k on both sides, we conclude that with probability at least $1 - \delta_k$,

$$\theta - 1/2^{k+2} \leq \theta^{(k)} + \arccos(\sqrt{\hat{X}})/2^k \leq \theta + 1/2^{k+2} .$$

Subtracting $1/2^{k+2}$ from all sides and recalling the definition of $\theta^{(k+1)}$ above, we conclude that $\theta^{(k+1)}$ is an underestimate of θ by at most $1/2^{k+1}$ as claimed. □

Continuing in this fashion up to $k = \overline{k} \triangleq \lceil \log_2 1/\epsilon \rceil$, we obtain an angle $\theta^{(\overline{k})}$ which underestimates θ by at most ϵ, with probability at least $1 - \sum_k \delta_k$. Suppose in each round k, we take $\delta_k \triangleq \delta 2^{k-\overline{k}-1}$, so that $\sum_k \delta_k \leq \delta$.

Furthermore, in any round k, we perform

$$2^k \cdot C \log 1/\delta_k = 2^k \cdot C \log 1/\delta + O(2^k \cdot C(\overline{k} + 1 - k))$$

queries to $U^{(k)}$ which amounts to as many queries to U. So the total number of queries made to U is

$$(1 + 2 + \cdots + 2^{\overline{k}}) C \log 1/\delta + \sum_{k=0}^{\overline{k}} O(2^k(\overline{k} + 1 - k)) = O(\log(1/\delta)/\epsilon)$$

as desired.

3. Looking Ahead

3.1. Rotation Learning in the Wild

The rotation learning problem can be thought of as a toy stand–in for a physical process that imprints a phase θ on a two–level system (a qubit, a pair of optical modes, or a two–dimensional invariant subspace inside a larger device).

In precision sensing (gravitational–wave interferometers, atomic clocks, Ramsey spectroscopy, and phase estimation in general), the central task is to learn a small θ as efficiently as possible. Real instruments like LIGO [**AAA+13**] do not literally implement the protocol we analyze here; for example, they inject *squeezed light* to reduce measurement noise rather than concatenating many coherent applications of the same unknown operation. Squeezing is notably more robust to realistic optical losses than schemes that try to amplify phase information solely by repeated coherent evolution or fragile entangled probes. Still, at the level of information flow, many metrology strategies can be idealized as:

$$\text{(prepare a known state)} \xrightarrow{\text{apply } U_\theta\text{, possibly with controls}} \text{(measure and update)}.$$

Our rotation–learning toy model captures precisely this prepare–evolve–measure loop. It allowed us to isolate two ingredients that matter for sample complexity: (i) how coherently we can *accumulate* phase information (e.g. by applying U multiple times or by clever controls), and (ii) how we post-process this information phase into a reliable estimate of the unknown quantum object. The formalism developed in this book will vastly generalize this example and its strategy.

3.2. Extensions

Although we illustrated the bootstrap idea with a 2×2 rotation, in fact the same idea can be extended to learn *any* 2×2 unitary matrix in $O(1/\epsilon)$ total queries.

In fact, one can even extend this beyond 2 dimensions. What is needed is an appropriate generalization of the step where we estimated the bias of a coin toss to constant error $\epsilon_{\mathsf{crude}}$. The relevant ideas for doing this will be introduced later on in this lecture when we discuss **tomography**. When we move from 2 dimensions to d dimensions however, the crucial change is that the number of queries will now depend on d. The intuition is that a completely unknown unitary on a d–dimensional space has d^2 real parameters, so without additional assumptions, one should not expect to learn all of these parameters until the number of queries

scales with $O(d^2)$. Indeed, it was shown recently [**HKOT23**] that the optimal query complexity for learning an arbitrary unitary matrix in d dimensions in the above model is exactly d^2/ϵ, up to constant factors. The argument we presented above is really just a baby version of the argument in that work.

Unfortunately, in the settings we will be interested in, we will always think of d as scaling *exponentially* in the number of "particles" in the system. To avoid exponential scaling, we then need to posit additional *structure* and align the learning task with that structure. For example, one standard and "physically reasonable" choice of structure to assume is that the unknown unitary takes the form of $U = e^{-iH}$, where H is a **local Hamiltonian** on n qubits; we will define this in due time, but for now the intuition to keep in mind is that this Hamiltonian is described by a total number of free parameters that only scales *polynomially* in the number of particles. Under such structural assumptions, one can then hope to develop algorithms that scale much more efficiently – we will cover these in a later unit in this course.

As another preview for what is to come, note that one can consider other models of interaction. In the query model we considered in this lecture, we allowed arbitrary control, and our choice of experiments was adaptive over the different rounds of the learning protocol. One could further enhance this model by, for instance, performing m entangled experiments in parallel, expanding the relevant dimension from d to d^m. While this doesn't end up buying much for the unitary learning problem, in many other quantum learning settings this can make a big difference in the efficiency with which one can learn. In the other direction, one can also consider *weaker* models where, perhaps due to various practical constraints on the experimental apparatus like hardware limitations or noise, we cannot perform arbitrary control. The effect of such constraints on the ultimate efficiency with which we can learn quantum states is another central theme in these notes.

Stepping further back, the rotation–learning example isolates three ingredients that will organize the rest of the book: the *unknown object* (a state, unitary, channel, or Hamiltonian), the *access model* (how we may prepare inputs, interleave known controls, parallelize or reuse the device, and measure), and the *loss metric* (in this case, the "parameter error" with which we estimate θ). Throughout the course of these lectures, we will use these basic ingredients to develop the foundations of a general theory of quantum learning.