Electronic Colloquium on Computational Complexity, Report No. 106 (2006)

The Learnability of Quantum States

Scott Aaronson*
University of Waterloo

Abstract

Traditional quantum state tomography requires a number of measurements that grows exponentially with the number of qubits n. But using ideas from computational learning theory, we show that "for most practical purposes" one can learn a state using a number of measurements that grows only *linearly* with n. Besides possible implications for experimental physics, our learning theorem has two applications to quantum computing: first, a new simulation of quantum one-way protocols, and second, the use of trusted classical advice to verify untrusted quantum advice.

1 Introduction

Suppose we have a physical process that produces a quantum state. By applying the process repeatedly, we can prepare as many copies of the state as we want, and can then measure each copy in a basis of our choice. The goal is to learn an approximate description of the state by combining the various measurement outcomes.

This problem is called *quantum state tomography*, and interest in it is more than hypothetical: it is already an important task in experimental physics. To give some examples, tomography has been used to obtain a detailed picture of a chemical reaction (namely, the dissociation of I_2 molecules) [30]; to confirm the preparation of 3-photon entangled states [29]; to test controlled-NOT gates [27]; and to characterize photodetectors and other optical devices [16].

Physicists would like to scale up tomography to larger systems, in order to study the many-particle entangled states that arise (for example) in chemistry, condensed-matter physics, and quantum information. But there is a fundamental obstacle, which has thus far prevented complete tomography of even *four*-qubit states. This is that, to reconstruct an n-qubit state, one needs to measure a number of observables that grows exponentially in n: in particular like 4^n, the number of parameters in a $2^n \times 2^n$ density matrix. This exponentiality is certainly a practical problem, but to us it is a conceptual problem as well. For what it suggests is that learning an arbitrary quantum state of (say) a hundred spin-1/2 particles would take longer than the age of the universe, even for a being with unlimited computational power. This, in turn, raises the question of what one even *means* when talking about the "state" of such a system. For whatever else a quantum state might be, *at the least* it ought to be a hypothesis that encapsulates previous observations of a physical system and allows us to predict future observations.

Our purpose here is to propose a new resolution of this conundrum. The idea is that, to predict the outcomes of "most" measurements of a quantum state, it is not necessary to do full tomography on the state. It suffices instead to do what we call "pretty good tomography," requiring a number of measurements that grows only *linearly* with the number of qubits. In the remainder of this introduction, we first give a formal statement of our result, then answer objections to it, situate it in the context of earlier work, and discuss its implications for theory and experiment.

1.1 Statement of Result

Let ρ be an n-qubit mixed state: that is, a $2^n \times 2^n$ Hermitian positive semidefinite matrix with $\mathrm{Tr}(\rho) = 1$. By a *measurement* of ρ, we will mean a "two-outcome POVM": that is, a $2^n \times 2^n$ Hermitian matrix E

*Email: scott@scottaaronson.com. Supported by CIAR through the Institute for Quantum Computing.

ISSN 1433-8092

with eigenvalues in $[0,1]$. Such a measurement E *accepts* ρ with probability $\mathrm{Tr}(E\rho)$, and *rejects* ρ with probability $1-\mathrm{Tr}(E\rho)$.

Our goal will be to learn ρ. Our notion of "learning" here is purely operational: we want a procedure that, given a measurement E, estimates the acceptance probability $\mathrm{Tr}(E\rho)$. Of course, estimating $\mathrm{Tr}(E\rho)$ for *every* E is the same as estimating ρ itself, and we know this requires exponentially many measurements. So if we want to learn ρ using fewer measurements, then we will have to settle for some weaker success criterion. The criterion we adopt is that we should be able to estimate $\mathrm{Tr}(E\rho)$ for *most* measurements E. In other words, we assume there is some (possibly unknown) probability distribution $\mathcal{D}$ from which the measurements are drawn.[1] We are given a "training set" of measurements $E_1,\ldots,E_m$ drawn independently from $\mathcal{D}$, as well as the approximate values of $\mathrm{Tr}(E_i\rho)$ for $i\in\{1,\ldots,m\}$. Our goal is to estimate $\mathrm{Tr}(E\rho)$ for most E's drawn from $\mathcal{D}$, with high probability over the choice of training set.

We will show that this can be done using a number of training measurements m that grows only *linearly* with the number of qubits n, and inverse-polynomially with the relevant error parameters. Furthermore, the learning procedure that achieves this bound is the simplest one imaginable: it suffices to find *any* "hypothesis state" σ such that $\mathrm{Tr}(E_i\sigma)\approx\mathrm{Tr}(E_i\rho)$ for all i. Then with high probability that hypothesis will "generalize," in the sense that $\mathrm{Tr}(E\sigma)\approx\mathrm{Tr}(E\rho)$ for most E's drawn from $\mathcal{D}$. More precisely:

Theorem 1.1 *Let ρ be an n-qubit mixed state, let $\mathcal{D}$ be a distribution over two-outcome measurements of ρ, and let $\mathcal{E}=(E_1,\ldots,E_m)$ be a "training set" consisting of m measurements drawn independently from $\mathcal{D}$. Also, fix error parameters $\varepsilon,\eta,\gamma>0$ with $\gamma\varepsilon\geq 7\eta$. Call $\mathcal{E}$ a "good" training set if any hypothesis σ that satisfies*

$$|\mathrm{Tr}(E_i\sigma)-\mathrm{Tr}(E_i\rho)|\leq\eta$$

for all $E_i\in\mathcal{E}$, also satisfies

$$\Pr_{E\in\mathcal{D}}\left[|\mathrm{Tr}(E\sigma)-\mathrm{Tr}(E\rho)|>\gamma\right]\leq\varepsilon.$$

Then there exists a constant $K>0$ such that $\mathcal{E}$ is a good training set with probability at least $1-\delta$, provided that

$$m\geq\frac{K}{\gamma^2\varepsilon^2}\left(\frac{n}{\gamma^2\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon}+\log\frac{1}{\delta}\right).$$

1.2 Objections and Variations

Before proceeding further, it will be helpful to answer various objections that might be raised against Theorem 1.1. Along the way, we will also state two variations of the theorem.

Objection 1 *By changing the goal to "pretty good tomography," Theorem 1.1 dodges much of the quantum state tomography problem as ordinarily understood.*

Response. Yes, that is exactly what it does! The motivating idea is that one does not need to know the expectation values for *all* observables, only for most of the observables that will actually be measured. As an example, if we can only apply 1- and 2-qubit measurements, then the outcomes of 3-qubit measurements are irrelevant by assumption. As a less trivial example, suppose the measurement distribution $\mathcal{D}$ is uniformly random (i.e., is the Haar measure). Then even if our quantum system is "really" in some pure state $|\psi\rangle$, for reasonably large n it will be trillions of years before we happen upon a measurement that distinguishes $|\psi\rangle$ from the maximally mixed state. It follows that the maximally mixed state is perfectly adequate as an *explanatory hypothesis*, despite being far from $|\psi\rangle$ in the usual metrics such as trace distance.

Of course, even after one relaxes the goal in this way, it might still seem surprising that for *any* state ρ, and *any* distribution $\mathcal{D}$, a linear amount of tomographic data is sufficient to simulate most measurements drawn from $\mathcal{D}$. This is the content of Theorem 1.1.

Objection 2 *But to apply Theorem 1.1, one would need to know the measurement distribution $\mathcal{D}$, which one almost never does in practice.*

[1] $\mathcal{D}$ can also be a continuous probability measure; this will not affect any of our results.

Response. This is incorrect: all of our learning algorithms will be "distribution-free," in the sense that a single algorithm will work for any $\mathcal{D}$. What we *do* need to assume is that future measurements will be drawn from the same distribution as past ones—or colloquially, that "nothing will be on the test that wasn't covered in class." As a trivial example, if all of the training data involved measurements in the $|0\rangle, |1\rangle$ basis, then clearly we couldn't expect to simulate measurements in the $|+\rangle, |-\rangle$ basis.

Objection 3 *Theorem 1.1 is purely information-theoretic; as such, it says nothing about the computational complexity of finding a hypothesis state σ.*

Response. This is correct. Using semidefinite and convex programming techniques, one can implement any of our learning algorithms to run in time polynomial in the Hilbert space dimension, $N = 2^n$. This might be fine if n is at most 10 or so; note that "measurement complexity," and not computational complexity, has almost always been the limiting factor in real experiments. But of course such a running time is prohibitive for larger n.

Let us stress that exactly the same problem arises even in *classical* learning theory. For it follows from a celebrated result of Goldreich, Goldwasser, and Micali [19] that, if finding a classical hypothesis consistent with data can always be done in polynomial time, then there are no cryptographic one-way functions. Using the same techniques, one can show that, if finding a *quantum* hypothesis consistent with data can always be done in *quantum* polynomial time, then there are no (classical) one-way functions secure against quantum attack. The only difference is that, while finding a classical hypothesis consistent with data is an NP search problem,[2] finding a quantum hypothesis is a QMA search problem.

A fundamental question left open by this paper is whether there are nontrivial special cases of the quantum learning problem that can be solved, not only with a linear number of measurements, but also with a polynomial amount of computation.

Objection 4 *The dependence on the error parameters γ and ε in Theorem 1.1 looks terrible.*

Response. Indeed, no one would pretend that performing $\sim \frac{1}{\gamma^4 \varepsilon^4}$ measurements is practical for reasonable γ and ε. Fortunately, we can improve the dependence on γ and ε quite substantially, at the cost of increasing the dependence on n from linear to $n \log^2 n$.

Theorem 1.2 *The bound in Theorem 1.1 can be replaced by*

$$m \geq \frac{K}{\varepsilon} \left(\frac{n}{(\gamma - \eta)^2} \log^2 \frac{n}{(\gamma - \eta)\varepsilon} + \log \frac{1}{\delta} \right)$$

for all $\varepsilon, \eta, \gamma > 0$ with $\gamma > \eta$.

In Section 4, we will show that the dependence on γ and ε in Theorem 1.2 is close to optimal.

Objection 5 *To estimate the measurement probabilities $\mathrm{Tr}(E_i \rho)$, one needs the ability to prepare multiple copies of ρ.*

Response. This is less an objection to Theorem 1.1 than to quantum mechanics itself! If one has only one copy of ρ, then Holevo's Theorem [20] immediately implies that not even "pretty good tomography" is possible.

Objection 6 *Even with unlimited copies of ρ, one could never be certain that the condition of Theorem 1.1 was satisfied (i.e., that $|\mathrm{Tr}(E_i \sigma) - \mathrm{Tr}(E_i \rho)| \leq \eta$ for every i).*

[2] Interestingly, in the "representation-independent" setting (where the output hypothesis can be an arbitrary Boolean circuit), this problem is *not* known to be NP-complete.

Response. This is correct, but there is no need for certainty. For suppose we apply each measurement E_i to $\Theta\left(\frac{\log m}{\eta^2}\right)$ copies of ρ. Then by a large deviation bound, with overwhelming probability we will obtain real numbers $p_1, \ldots, p_m$ such that $|p_i - \mathrm{Tr}(E_i\rho)| \leq \eta/2$ for every i. So if we want to find a hypothesis state σ such that $|\mathrm{Tr}(E_i\sigma) - \mathrm{Tr}(E_i\rho)| \leq \eta$ for every i, then it suffices to find a σ such that $|p_i - \mathrm{Tr}(E_i\sigma)| \leq \eta/2$ for every i. Certainly such a σ exists, for take $\sigma = \rho$.

Objection 7 *But what if one can only apply each measurement once, rather than multiple times? In that case, the above estimation strategy no longer works.*

Response. In Section 3, we will prove a learning theorem that applies directly to this "measure-once" scenario. The disadvantage is that the upper bound on measurement complexity will increase from $\sim 1/\left(\gamma^4\varepsilon^4\right)$ to $\sim 1/\left(\gamma^8\varepsilon^4\right)$.

Theorem 1.3 *Let ρ be an n-qubit state, let $\mathcal{D}$ be a distribution over two-outcome measurements, and let $\mathcal{E} = (E_1, \ldots, E_m)$ consist of m measurements drawn independently from $\mathcal{D}$. Suppose we are given bits $B = (b_1, \ldots, b_m)$, where each b_i is 1 with independent probability $\mathrm{Tr}(E_i\rho)$ and 0 with probability $1 - \mathrm{Tr}(E_i\rho)$. Suppose also that we choose a hypothesis state σ to minimize the quadratic functional $\sum_{i=1}^m \left(\mathrm{Tr}(E_i\sigma) - b_i\right)^2$. Then there exists a positive constant K such that*

$$\Pr_{E\in\mathcal{D}}\left[|\mathrm{Tr}(E\sigma) - \mathrm{Tr}(E\rho)| > \gamma\right] \leq \varepsilon$$

with probability at least $1-\delta$ over $\mathcal{E}$ and B, provided that

$$m \geq \frac{K}{\gamma^4\varepsilon^2}\left(\frac{n}{\gamma^4\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\right).$$

1.3 Related Work

This paper builds on two research areas—computational learning theory and quantum information theory—in order to say something about a third area: quantum state estimation. Since many readers are probably unfamiliar with at least one of these areas, let us discuss them in turn.

Computational learning theory can be understood as a modern response to David Hume's Problem of Induction: "if an ornithologist sees 500 crows and all of them are black, why does that provide *any grounds at all* for expecting the 501^{st} crow to be black? After all, the hypothesis that the 501^{st} crow will be white would seem equally compatible with evidence." The answer, from a learning theory perspective, is that in practice one always restricts attention to some class $\mathcal{C}$ of hypotheses that is vastly smaller than the class of logically conceivable hypotheses. So the real question is not "is induction possible?," but rather "what properties does the class $\mathcal{C}$ have to satisfy for induction to be possible?"

In a seminal 1984 paper, Valiant [32] showed that if $\mathcal{C}$ is finite, then any hypothesis that agrees with $O(\log|\mathcal{C}|)$ randomly-chosen data points will probably agree with most future data points as well. Subsequently Blumer et al. [12] showed that even if $\mathcal{C}$ is infinite, one can upper-bound the number of data points needed for learning in terms of a combinatorial parameter of $\mathcal{C}$ called the VC (Vapnik-Chervonenkis) dimension. Unfortunately, both of these results apply only to Boolean hypothesis classes. So to prove our quantum learning theorem, we will need more powerful results due to Anthony and Bartlett [9] and Bartlett and Long [10], which upper-bound the number of data points needed to learn *real*-valued hypothesis classes.

Besides these learning theory results, we will also need a result of Nayak [25] in quantum information theory. What Nayak showed is that, if we want to encode k bits into an n-qubit quantum state, in such a way that any one bit can later be retrieved with error probability at most p, then we must have $n \geq (1 - H(p))\,k$, where H is the binary entropy function. This tightened an earlier result of Ambainis et al. [8], that $\Omega(k/\log k)$ qubits are necessary. Although the motivation for these results came from coding and automata problems, one can interpret the results much more broadly, as upper-bounding the number of *independently-accessible degrees of freedom in a quantum state*. Nayak's proof works by reduction to Holevo's Theorem [20], which states that an n-qubit quantum channel can reliably transmit at most n classical bits. Holevo's Theorem can in turn be seen as a fundamental consequence of the linearity of quantum mechanics.

Physicists have been interested in quantum state estimation since at least the 1950's (see [28] for an excellent introduction to the area). For practical reasons, they have been particularly concerned with minimizing the number of measurements. However, almost all literature on the subject restricts attention to low-dimensional Hilbert spaces (say, 2 or 3 qubits), taking for granted that the number of measurements will increase exponentially with the number of qubits. There *is* a substantial literature on how to estimate a quantum state from incomplete measurement results, using the Maximum-Likelihood or Maximum-Entropy principles—see Bužek [14] or Bužek et al. [15] for example. But while some of this work has offered numerical evidence that few measurements seem to suffice in practice, so far as we know none of it has considered asymptotic complexity.

1.4 Implications

For us the main implication of our learning theorem is conceptual. Over the past few decades, the field of computational learning theory has managed to delineate which hypothesis classes are "reasonable," in the sense that any hypothesis compatible with a modest number of data points will probably predict most future data as well. Our result shows that quantum states, considered as a hypothesis class, are "reasonable" in the learning theory sense. Were this *not* the case, it would presumably strengthen the view of quantum computing skeptics [18, 24] that quantum states are "inherently extravagant" objects, which will need to be discarded as our knowledge of physics expands.[3] Instead we have shown that, while the "effective dimension" of an n-qubit Hilbert space *appears* to be exponential in n, in the sense that is relevant for approximate learning and prediction this appearance is illusory.

Beyond establishing this conceptual point, we believe our learning theorem might be of practical use in quantum state estimation, since it provides experimenters with an explicit upper bound on the number of measurements needed to "learn" a quantum state with respect to any probability measure over observables. Even if our actual upper bounds are too weak, or are otherwise not what is needed, we hope the mere *fact* that this sort of learning is possible will serve as a spur to further research. As an analogy, the classical learning theorems of Valiant [32] and others have had a large influence on neural networks, computer vision, and other fields,[4] though sometimes the influence has been subtle and indirect.

We turn now to a more immediate application of our learning theorem: solving open problems in quantum computing and information.

The first problem concerns *quantum one-way communication complexity*. In this subject we consider a sender, Alice, and a receiver, Bob, who hold inputs x and y respectively. We then ask the following question: assuming the best communication protocol and the worst (x,y) pair, how many bits must Alice send to Bob, for Bob to be able to evaluate some joint function $f(x,y)$? Note that there is no back-communication from Bob to Alice.

Let $\mathrm{D}^1(f)$, $\mathrm{R}^1(f)$, and $\mathrm{Q}^1(f)$ be the number of bits that Alice needs to send, if her message to Bob is deterministic, randomized, or quantum respectively.[5] Then we are interested in what relationships hold between these measures. Certainly $\mathrm{D}^1(f) \geq \mathrm{R}^1(f) \geq \mathrm{Q}^1(f)$ for any f. A few years ago, Aaronson [2] showed that for any Boolean function f (partial or total), $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\log \mathrm{Q}^1(f)\right)$, where M is the number of bits in Bob's input. Intuitively, this means that if Bob's input is small, then quantum communication provides at most a small advantage over classical communication.

An obvious question about Aaronson's upper bound, but one that resisted attack, is whether the $\log \mathrm{Q}^1(f)$ factor can be removed. In Section 5 we will show that it *can* be removed, *provided* we replace $\mathrm{D}^1(f)$ by $\mathrm{R}^1(f)$:

Theorem 1.4 *For any Boolean function f (partial or total), $\mathrm{R}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$, where M is the length of Bob's input.*

[3] Or at least, it would suggest that the "operationally meaningful" quantum states comprise only a tiny portion of Hilbert space. See Aaronson [1] for an attempt to make this notion precise.

[4] According to Google Scholar, Valiant's paper has been cited 1772 times, with a large fraction of the citations coming from practitioners.

[5] Here the superscript '1' denotes one-way communication. Also, if Alice's message is randomized or quantum, then Bob can output a wrong answer with $\frac{1}{3}$ probability.

The proof of Theorem 1.4 will rely on our learning theorem in an intuitively appealing way. Basically, Alice will send some randomly-chosen "training inputs," which Bob will then use to learn a "pretty good description" of the quantum state that Alice would have sent him in the quantum protocol.

The second problem concerns *quantum advice*. Intuitively, a "quantum advice state" is a state that might be extremely hard to prepare, but that would dramatically extend our computational powers *if* we were somehow given it to use as the initial state of a quantum computer. To formalize this concept, Nishimura and Yamakami [26] defined a complexity class called BQP/qpoly, or BQP with polynomial-size quantum advice.[6] This class consists of all problems that are solvable in polynomial time by a quantum computer, with help from a polynomial-size quantum advice state $|\psi_n\rangle$ that depends only on the input length n but can otherwise be arbitrary. How powerful is this class? We know that BQP/qpoly strictly contains BQP, since the advice state $|\psi_n\rangle$ might encode (for example) whether or not the n^{th} Turing machine halts. So the interesting question is whether BQP/qpoly strictly contains BQP/poly, or BQP with polynomial-size *classical* advice. In other words, can quantum advice ever be exponentially more powerful than classical advice? Perhaps not surprisingly, the answer is unknown.[7]

Aaronson [2] at least showed that BQP/qpoly is contained in another class called PostBQP/poly. Intuitively, this means that quantum advice is not "infinitely" powerful: one can always replace it by classical advice, provided one uses exponentially more computation time. A natural question, but one that remained open for several years, is whether Aaronson's result can be improved to BQP/qpoly $\subseteq$ QMA/poly. Here QMA, or Quantum Merlin-Arthur, is the class of problems for which a 'yes' answer admits a polynomial-size quantum proof. Loosely speaking, such a containment would mean that trusted quantum advice can always be replaced by *untrusted* quantum advice together with trusted *classical* advice—or in other words, that trusted classical advice can be used to *verify* untrusted quantum advice.

Though we are unable to prove that BQP/qpoly $\subseteq$ QMA/poly, in Section 6 we will do something close. We will prove that HeurBQP/qpoly $\subseteq$ HeurQMA/poly, where HeurBQP/qpoly and HeurQMA/poly are the *heuristic* versions of BQP/qpoly and QMA/poly respectively—that is, the versions where we only want to succeed on *most* inputs rather than all of them. This containment is actually a consequence of a stronger result:

Theorem 1.5 HeurBQP/qpoly = HeurYQP/poly.

Here YQP, or "Yoda Quantum Polynomial-Time," is a new subclass of QMA that we define and explore in Section 6.2, and which might be of independent interest. Roughly speaking, YQP is like BQP/qpoly, except that the quantum advice cannot be trusted.

To prove Theorem 1.5, the key idea is to provide the quantum algorithm with randomly-chosen "test inputs," which it can use to check an alleged quantum advice state prior to using it. By appealing to our learning theorem, we will show that if a quantum advice state yields the right answers on the test inputs, then with high probability it will yield the right answers on most other inputs as well. An obvious difficulty is that the measurements intended to *test* the quantum advice might also *destroy* it. The main technical contribution of Section 6 is to resolve this difficulty, by means of a so-called "Witness Protection Lemma."

2 The Measurement Complexity of Quantum Learning

In this section we prove Theorems 1.1 and 1.2. To do so, we first review results from computational learning theory, which upper-bound the number of data points needed to learn a hypothesis in terms of the "dimension" of the underlying hypothesis class. We then use a result of Nayak [25] to upper-bound the dimension of the class of n-qubit mixed states.

2.1 Learning Probabilistic Concepts

The prototype of the sort of learning theory result we need is the so-called "Occam's Razor Theorem" of Blumer et al. [12], which is stated in terms of a parameter called the VC dimension. However, Blumer et

[6] Here BQP is Bounded-Error Quantum Polynomial-Time: informally, the class of problems that are solvable efficiently by a quantum computer.

[7] Though see Aaronson and Kuperberg [6] for a "quantum oracle separation" and other partial results.

al.'s result does not suffice for our purpose, since it deals with *Boolean* concepts, which map each element of an underlying sample space to $\{0,1\}$. By contrast, we are interested in "probabilistic concepts"—called *p-concepts* by Kearns and Schapire [21]—which map each measurement E to a real number $\mathrm{Tr}(E\rho) \in [0,1]$.

Generalizing from Boolean concepts to p-concepts is not nearly as straightforward as one might think. Fortunately, various authors [7, 9, 10, 11, 21] have already done most of the work for us. Results due to Anthony and Bartlett [9] and to Bartlett and Long [10] will be particularly useful. To state these results, we first need some definitions. Let $\mathcal{S}$ be a finite or infinite set called the *sample space*. Then a *p-concept over* $\mathcal{S}$ is a function $F : \mathcal{S} \rightarrow [0,1]$, and a *p-concept class over* $\mathcal{S}$ is a set of p-concepts over $\mathcal{S}$. Kearns and Schapire [21] proposed a measure of the complexity of p-concept classes, called the *fat-shattering dimension*.

Definition 2.1 *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\gamma > 0$ be a real number. We say a set $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is γ-fat-shattered by $\mathcal{C}$ if there exist real numbers $\alpha_1, \ldots, \alpha_k$ such that for all $B \subseteq \{1, \ldots, k\}$, there exists a p-concept $F \in \mathcal{C}$ such that for all $i \in \{1, \ldots, k\}$,*

(i) if $i \notin B$ then $F(s_i) \leq \alpha_i - \gamma$, and

(ii) if $i \in B$ then $F(s_i) \geq \alpha_i + \gamma$.

Then the γ-fat-shattering dimension of $\mathcal{C}$, or $\mathrm{fat}_{\mathcal{C}}(\gamma)$, is the maximum k such that some $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is γ-fat-shattered by $\mathcal{C}$. (If no finite set is γ-fat-shattered by $\mathcal{C}$, then $\mathrm{fat}_{\mathcal{C}}(\gamma) = \infty$.)

We can now state the result of Anthony and Bartlett.

Theorem 2.2 (Anthony and Bartlett [9]) *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Fix an element $F \in \mathcal{C}$, as well as error parameters $\varepsilon, \eta, \gamma > 0$ with $\gamma > \eta$. Suppose we draw m samples $X = (x_1, \ldots, x_m)$ independently according to $\mathcal{D}$, and then choose any hypothesis $H \in \mathcal{C}$ such that $|H(x) - F(x)| \leq \eta$ for all $x \in X$. Then there exists a positive constant K such that*

$$\Pr_{x \in \mathcal{D}}[|H(x) - F(x)| > \gamma] \leq \varepsilon$$

with probability at least $1 - \delta$ over X, provided that

$$m \geq \frac{K}{\varepsilon}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma - \eta}{8}\right)\log^2\left(\frac{\mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma-\eta}{8}\right)}{(\gamma-\eta)\,\varepsilon}\right) + \log\frac{1}{\delta}\right).$$

Notice that in Theorem 2.2, the dependence on the fat-shattering dimension is superlinear. We would like to reduce the dependence to linear, at least when η is sufficiently small. We can do so using the following result of Bartlett and Long.[8]

Theorem 2.3 (Bartlett and Long [10]) *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Fix a p-concept $F : \mathcal{S} \rightarrow [0,1]$ (not necessarily in $\mathcal{C}$), as well as an error parameter $\alpha > 0$. Suppose we draw m samples $X = (x_1, \ldots, x_m)$ independently according to $\mathcal{D}$, and then choose any hypothesis $H \in \mathcal{C}$ such that $\sum_{i=1}^{m} |H(x_i) - F(x_i)|$ is minimized. Then there exists a positive constant K such that*

$$\mathop{\mathrm{EX}}_{x \in \mathcal{D}}[|H(x) - F(x)|] \leq \alpha + \inf_{C \in \mathcal{C}} \mathop{\mathrm{EX}}_{x \in \mathcal{D}}[|C(x) - F(x)|]$$

with probability at least $1 - \delta$ over X, provided that

$$m \geq \frac{K}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\alpha}{5}\right)\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right).$$

Theorem 2.3 has the following corollary.

[8] The result we state is a special case of Bartlett and Long's Theorem 20, where the function F to be learned is itself a member of the hypothesis class $\mathcal{C}$.

Corollary 2.4 *In the statement of Theorem 2.2, suppose $\gamma\varepsilon \geq 7\eta$. Then the bound on m can be replaced by*

$$m \geq \frac{K}{\gamma^2\varepsilon^2}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma\varepsilon}{35}\right)\log^2\frac{1}{\gamma\varepsilon}+\log\frac{1}{\delta}\right).$$

Proof. Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Then let $\mathcal{C}^*$ be the class of p-concepts $G : \mathcal{S} \to [0,1]$ for which there exists an $F \in \mathcal{C}$ such that $|G(x) - F(x)| \leq \eta$ for all $x \in \mathcal{S}$. Also, fix a p-concept $F \in \mathcal{C}$. Suppose we draw m samples $X = (x_1, \ldots, x_m)$ independently according to $\mathcal{D}$, and then choose any hypothesis $H \in \mathcal{C}$ such that $|H(x) - F(x)| \leq \eta$ for all $x \in X$. Then there exists a $G \in \mathcal{C}^*$ such that $G(x) = H(x)$ for all $x \in X$. This G is simply obtained by setting $G(x) := H(x)$ if $x \in X$ and $G(x) := F(x)$ otherwise.

So by Theorem 2.3, provided that

$$m \geq \frac{K}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\alpha}{5}\right)\log^2\frac{1}{\alpha}+\log\frac{1}{\delta}\right),$$

we have

$$\mathop{\mathrm{EX}}_{x\in\mathcal{D}}[|H(x) - G(x)|] \leq \alpha + \inf_{C\in\mathcal{C}^*}\mathop{\mathrm{EX}}_{x\in\mathcal{D}}[|C(x) - G(x)|] = \alpha$$

with probability at least $1-\delta$ over X. Here we have used the fact that $G \in \mathcal{C}^*$ and hence

$$\inf_{C\in\mathcal{C}^*}\mathop{\mathrm{EX}}_{x\in\mathcal{D}}[|C(x) - G(x)|] = 0.$$

Setting $\alpha := \frac{6\gamma}{7}\varepsilon$, this implies by Markov's inequality that

$$\Pr_{x\in\mathcal{D}}\left[|H(x) - G(x)| > \frac{6\gamma}{7}\right] \leq \varepsilon,$$

and therefore

$$\Pr_{x\in\mathcal{D}}\left[|H(x) - F(x)| > \frac{6\gamma}{7} + \eta\right] \leq \varepsilon.$$

Since $\eta \leq \frac{\gamma\varepsilon}{7} \leq \frac{\gamma}{7}$, the above implies that

$$\Pr_{x\in\mathcal{D}}[|H(x) - F(x)| > \gamma] \leq \varepsilon$$

as desired.

Next we claim that $\mathrm{fat}_{\mathcal{C}^*}(\alpha) \leq \mathrm{fat}_{\mathcal{C}}(\alpha - \eta)$. The reason is simply that, if a given set α-fat-shatters $\mathcal{C}^*$, then it must also $(\alpha-\eta)$-fat-shatter $\mathcal{C}$ by the triangle inequality.

Putting it all together, we have

$$\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\alpha}{5}\right) \leq \mathrm{fat}_{\mathcal{C}}\left(\frac{\alpha}{5}-\eta\right) \leq \mathrm{fat}_{\mathcal{C}}\left(\frac{6\gamma\varepsilon/7}{5}-\frac{\gamma\varepsilon}{7}\right) = \mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma\varepsilon}{35}\right),$$

and hence

$$\begin{aligned} m &\geq \frac{K}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma\varepsilon}{35}\right)\log^2\frac{1}{\alpha}+\log\frac{1}{\delta}\right) \\ &= \frac{K}{(6\gamma\varepsilon/7)^2}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\gamma\varepsilon}{35}\right)\log^2\frac{1}{6\gamma\varepsilon/7}+\log\frac{1}{\delta}\right) \end{aligned}$$

samples suffice. ■

2.2 Learning Quantum States

We now turn to the problem of learning a quantum state. Let $\mathcal{S}$ be the set of two-outcome measurements on n qubits. Also, given an n-qubit mixed state ρ, let $F_\rho : \mathcal{S} \rightarrow [0,1]$ be the p-concept defined by $F_\rho(E) = \mathrm{Tr}(E\rho)$, and let $\mathcal{C}_n = \{F_\rho\}_\rho$ be the class of all such F_ρ's. Then to apply Theorems 2.2 and 2.3, all we need to do is upper-bound $\mathrm{fat}_{\mathcal{C}_n}(\gamma)$ in terms of n and γ. We will do so using a result of Nayak [25], which upper-bounds the number of classical bits that can be "encoded" into n qubits.

Theorem 2.5 (Nayak [25]) *Let k and n be positive integers with $k > n$. For all k-bit strings $y = y_1 \cdots y_k$, let ρ_y be an n-qubit mixed state that "encodes" y. Suppose there exist two-outcome measurements $E_1, \ldots, E_k$ such that for all $y \in \{0,1\}^k$ and $i \in \{1, \ldots, k\}$,*

(i) if $y_i = 0$ then $\mathrm{Tr}(E_i \rho_y) \leq p$, and

(ii) if $y_i = 1$ then $\mathrm{Tr}(E_i \rho_y) \geq 1 - p$.

Then $n \geq (1 - H(p))\, k$, where H is the binary entropy function.

Theorem 2.5 has the following easy generalization.

Theorem 2.6 *Let k, n, and $\{\rho_y\}$ be as in Theorem 2.5. Suppose there exist measurements $E_1, \ldots, E_k$, as well as real numbers $\alpha_1, \ldots, \alpha_k$, such that for all $y \in \{0,1\}^k$ and $i \in \{1, \ldots, k\}$,*

(i) if $y_i = 0$ then $\mathrm{Tr}(E_i \rho_y) \leq \alpha_i - \gamma$, and

(ii) if $y_i = 1$ then $\mathrm{Tr}(E_i \rho_y) \geq \alpha_i + \gamma$.

Then $n/\gamma^2 = \Omega(k)$.

Proof. Suppose there exists such an encoding scheme with $n/\gamma^2 = o(k)$. Then consider an amplified scheme, where each string $y \in \{0,1\}^k$ is encoded by the tensor product state $\rho_y^{\otimes \ell}$. Here we set $\ell := \lceil c/\gamma^2 \rceil$ for some $c > 0$. Also, for all $i \in \{1, \ldots, k\}$, let E_i^* be an amplified measurement that applies E_i to each of the ℓ copies of ρ_y, and accepts if and only if at least $\alpha_i \ell$ of the E_i's do. Then provided we choose c sufficiently large, it is easy to show by a Chernoff bound that for all y and i,

(i) if $y_i = 0$ then $\mathrm{Tr}\left(E_i^* \rho_y^{\otimes \ell}\right) \leq \frac{1}{3}$, and

(ii) if $y_i = 1$ then $\mathrm{Tr}\left(E_i^* \rho_y^{\otimes \ell}\right) \geq \frac{2}{3}$.

So to avoid contradicting Theorem 2.5, we need $n\ell \geq \left(1 - H\left(\frac{1}{3}\right)\right) k$. But this implies that $n/\gamma^2 = \Omega(k)$.[9] ■

If we interpret k as the size of a fat-shattered subset of $\mathcal{S}$, then Theorem 2.6 immediately yields the following upper bound on fat-shattering dimension.

Corollary 2.7 *For all $\gamma > 0$ and n, we have $\mathrm{fat}_{\mathcal{C}_n}(\gamma) = O\left(n/\gamma^2\right)$.*

Combining Corollary 2.4 with Corollary 2.7, we find that if $\gamma\varepsilon \geq 7\eta$, then it suffices to use

$$\begin{aligned} m &= \left\lceil \frac{K}{\gamma^2 \varepsilon^2} \left(\mathrm{fat}_{\mathcal{C}_n}\left(\frac{\gamma\varepsilon}{35}\right) \log^2 \frac{1}{\gamma\varepsilon} + \log \frac{1}{\delta} \right) \right\rceil \\ &= O\left(\frac{1}{\gamma^2 \varepsilon^2} \left(\frac{n}{\gamma^2 \varepsilon^2} \log^2 \frac{1}{\gamma\varepsilon} + \log \frac{1}{\delta} \right) \right) \end{aligned}$$

[9] If we care about optimizing the constant under the $\Omega(k)$, then we are better off avoiding amplication and instead proving Theorem 2.6 directly using the techniques of Nayak [25]. Doing so, we obtain $n/\gamma^2 \geq 2k/\ln 2$.

measurements. Likewise, combining Theorem 2.2 with Corollary 2.7, we find that if $\gamma > \eta$, then it suffices to use

$$\begin{aligned} m &= \left\lceil \frac{K}{\varepsilon} \left(\mathrm{fat}_{\mathcal{C}_n} \left(\frac{\gamma - \eta}{8} \right) \log^2 \left(\frac{\mathrm{fat}_{\mathcal{C}_n} \left(\frac{\gamma - \eta}{8} \right)}{(\gamma - \eta)\varepsilon} \right) + \log \frac{1}{\delta} \right) \right\rceil \\ &= O\left(\frac{1}{\varepsilon} \left(\frac{n}{(\gamma - \eta)^2} \log^2 \frac{n}{(\gamma - \eta)\varepsilon} + \log \frac{1}{\delta} \right) \right) \end{aligned}$$

measurements. This completes the proofs of Theorems 1.1 and 1.2 respectively.

In Section 4, we will prove a corresponding lower bound: namely, that there exists a measurement distribution $\mathcal{D}$ for which

$$m = \Omega\left(\frac{1}{\varepsilon} \left(\frac{n}{\gamma^2} + \log \frac{1}{\delta} \right) \right)$$

measurements are necessary. In particular, if we suppose η is negligible, then Theorem 1.1 is tight in its dependence on n, while Theorem 1.2 is tight up to a multiplicative factor of $\log^2(n/\gamma)$.

3 Learning from Measurement Results

In Section 2 we considered a model where for each measurement E, the learner is told the approximate value of $\mathrm{Tr}(E\rho)$. This model will suffice for our applications to quantum computing and information. But for other applications, it might be natural to ask what happens if we instead assume that for each E, the learner is merely given a measurement outcome: that is, a bit that is 1 with probability $\mathrm{Tr}(E\rho)$ and 0 with probability $1 - \mathrm{Tr}(E\rho)$. Of course, if the learner were given many such measurement outcomes for the same E, it could form an estimate of $\mathrm{Tr}(E\rho)$. But we are assuming that for each E, the learner only receives one measurement outcome.

We will show that, even in this seemingly weak model, an n-qubit quantum state can still be learned using $O(n)$ measurements, although the dependence on the parameters γ and ε will worsen.

The general task we are considering—that of learning a p-concept given only samples from its associated probability distribution—is called *learning in the p-concept model*. The task was first studied in the early 1990's by Kearns and Schapire [21], who left as an open question whether it can always be done if the fat-shattering dimension is finite. Alon et al. [7] answered this question affirmatively in a breakthrough a few years later. Unfortunately, Alon et al. never worked out the actual complexity bound implied by their result, and to the best of our knowledge no one else did either. Thus, our first task will be to fill this rather large gap in the literature. We will do so using Theorem 2.3, which was proven by Bartlett and Long [10] building on ideas of Alon et al.

Theorem 3.1 *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Fix a p-concept $F \in \mathcal{C}$, as well as error parameters $\varepsilon, \gamma > 0$. Suppose we are given m samples $X = (x_1, \ldots, x_m)$ drawn independently from $\mathcal{D}$, as well as bits $B = (b_1, \ldots, b_m)$ such that each b_i is 1 with independent probability $F(x_i)$. Suppose also that we choose a hypothesis $H \in \mathcal{C}$ to minimize the quadratic functional $\sum_{i=1}^m (H(x_i) - b_i)^2$. Then there exists a positive constant K such that*

$$\Pr_{x \in \mathcal{D}} [|H(x) - F(x)| > \gamma] \leq \varepsilon$$

with probability at least $1 - \delta$ over X and B, provided that

$$m \geq \frac{K}{\gamma^4 \varepsilon^2} \left(\mathrm{fat}_{\mathcal{C}} \left(\frac{\gamma^2 \varepsilon}{10} \right) \log^2 \frac{1}{\gamma \varepsilon} + \log \frac{1}{\delta} \right).$$

Proof. Let $\mathcal{D}^*$ be the distribution over $(x, b) \in \mathcal{S} \times \{0, 1\}$ obtained by first drawing x from $\mathcal{D}$, and then setting $b = 1$ with probability $F(x)$ and $b = 0$ with probability $1 - F(x)$. Then we can imagine that each (x_i, b_i) was drawn from $\mathcal{D}^*$. Also, given a hypothesis $H \in \mathcal{C}$, let $H^* : \mathcal{S} \times \{0, 1\} \rightarrow [0, 1]$ be the quadratic

loss function defined by $H^*(x,0) = H(x)^2$ and $H^*(x,1) = (1-H(x))^2$. Then let $\mathcal{C}^*$ be the p-concept class consisting of H^* for all $H \in \mathcal{C}$.

Call $H^* \in \mathcal{C}^*$ an "α-good" function if

$$\mathop{\mathrm{EX}}_{(x,b)\in\mathcal{D}^*}[H^*(x,b)] \leq \alpha + \inf_{C^*\in\mathcal{C}^*} \mathop{\mathrm{EX}}_{(x,b)\in\mathcal{D}^*}[C^*(x,b)].$$

Notice that

$$\begin{aligned}\mathop{\mathrm{EX}}_{(x,b)\in\mathcal{D}^*}[H^*(x,b)] &= \mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(1-F(x))H(x)^2 + F(x)(1-H(x))^2\right] \\ &= \mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(H(x)-F(x))^2 + F(x) - F(x)^2\right].\end{aligned}$$

Therefore, if H^* is α-good then

$$\mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(H(x)-F(x))^2\right] \leq \alpha + \inf_{C\in\mathcal{C}} \mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(C(x)-F(x))^2\right].$$

Since

$$\inf_{C\in\mathcal{C}} \mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(C(x)-F(x))^2\right] = 0,$$

this implies in particular that

$$\mathop{\mathrm{EX}}_{x\in\mathcal{D}}\left[(H(x)-F(x))^2\right] \leq \alpha.$$

If we set $\alpha := \gamma^2\varepsilon$, then the above implies by Markov's inequality that

$$\Pr_{x\in\mathcal{D}}[|H(x)-F(x)| > \gamma] \leq \varepsilon$$

as desired.

Now suppose we are given m samples $(x_1,b_1),\ldots,(x_m,b_m)$ drawn independently from $\mathcal{D}^*$. Also, let $Z(x,b) = 0$ be the identically-zero function. Then Theorem 2.3 implies that, if we choose $H^* \in \mathcal{C}^*$ to minimize

$$\sum_{i=1}^m |H^*(x_i,b_i) - Z(x_i,b_i)| = \sum_{i=1}^m H^*(x_i,b_i) = \sum_{i=1}^m (H(x_i)-b_i)^2,$$

then H^* will be α-good with probability at least $1-\delta$, provided that

$$\begin{aligned}m &= O\left(\frac{1}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\alpha}{5}\right)\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right) \\ &= O\left(\frac{1}{\gamma^4\varepsilon^2}\left(\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\gamma^2\varepsilon}{5}\right)\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\right)\right)\end{aligned}$$

Finally, we claim that $\mathrm{fat}_{\mathcal{C}^*}(\eta) \leq 2\,\mathrm{fat}_{\mathcal{C}}(\eta/2)$ for all $\eta > 0$. To see this, let $\mathcal{C}_0$ be the p-concept class consisting of $H(x)^2$ for all $H \in \mathcal{C}$, and let $\mathcal{C}_1$ be the class consisting of $(1-H(x))^2$ for all $H \in \mathcal{C}$. Then clearly

$$\mathrm{fat}_{\mathcal{C}^*}(\eta) \leq \mathrm{fat}_{\mathcal{C}_0}(\eta) + \mathrm{fat}_{\mathcal{C}_1}(\eta).$$

Also, if $\left|H_1(x)^2 - H_2(x)^2\right| \geq \eta$, then $|H_1(x) - H_2(x)| \geq \eta/2$. Hence $\mathrm{fat}_{\mathcal{C}_0}(\eta) \leq \mathrm{fat}_{\mathcal{C}}(\eta/2)$, and similarly $\mathrm{fat}_{\mathcal{C}_1}(\eta) \leq \mathrm{fat}_{\mathcal{C}}(\eta/2)$. ■

We can now prove Theorem 1.3: that

$$m = O\left(\frac{1}{\gamma^4\varepsilon^2}\left(\frac{n}{\gamma^4\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\right)\right)$$

measurements suffice to learn an n-qubit quantum state in the p-concept model.

Proof of Theorem 1.3. As in Section 2, let $\mathcal{C}_n$ be the class of functions $f_\rho : \mathcal{S} \rightarrow [0,1]$ defined by $f_\rho(E) = \mathrm{Tr}(E\rho)$. Then the theorem follows immediately from Theorem 3.1, together with the fact that

$$\mathrm{fat}_{\mathcal{C}_n}\left(\frac{\gamma^2\varepsilon}{10}\right) = O\left(\frac{n}{\left(\gamma^2\varepsilon/10\right)^2}\right) = O\left(\frac{n}{\gamma^4\varepsilon^2}\right)$$

by Corollary 2.7. ■

Given a state ρ, Theorem 1.3 upper-bounds the number of measurements needed to estimate the measurement probabilities $\mathrm{Tr}(E\rho)$. Can we do better if, instead of *estimating* the probabilities, we merely want to *predict* the outcomes themselves with nontrivial bias? In Appendix 9, we will prove an almost-tight variant of Theorem 1.3 that is optimized for this prediction task.

4 Lower Bounds

Having proved upper bounds on the measurement complexity of quantum learning, in this section we turn to proving lower bounds. Roughly speaking, we will show that there exists a measurement distribution $\mathcal{D}$ for which

$$m = \Omega\left(\frac{1}{\varepsilon}\left(\frac{n}{\gamma^2} + \log\frac{1}{\delta}\right)\right)$$

measurements are necessary. Also, in the model of Section 3—the model where each measurement is applied only once—we will show that

$$m = \Omega\left(\frac{1}{\varepsilon}\left(\frac{n}{\gamma^4} + \log\frac{1}{\delta}\right)\right)$$

measurements are necessary.[10] More formally:

Theorem 4.1 *Fix an integer $n > 0$ and error parameters $\varepsilon, \delta, \gamma \in (0,1)$. Then there exists a distribution $\mathcal{D}$ over n-qubit measurements for which the following holds.*

(i) Suppose

$$m = o\left(\frac{1}{\varepsilon}\left(\frac{n}{\gamma^2} + \log\frac{1}{\delta}\right)\right).$$

Then there is no learning algorithm that, given measurements $\mathcal{E} = (E_1, \ldots, E_m)$ drawn independently from $\mathcal{D}$, as well as real numbers $p_1, \ldots, p_m$ such that $|p_i - \mathrm{Tr}(E_i\rho)| \leq \gamma^2/n$ for all i, outputs a hypothesis state σ such that

$$\Pr_{E\in\mathcal{D}}\left[\left|\mathrm{Tr}(E\sigma) - \mathrm{Tr}(E\rho)\right| > \gamma\right] \leq \varepsilon$$

with probability at least $1-\delta$ over $\mathcal{E}$.

(ii) Suppose

$$m = o\left(\frac{1}{\varepsilon}\left(\frac{n}{\gamma^4} + \log\frac{1}{\delta}\right)\right).$$

Then there is no learning algorithm that, given measurements $\mathcal{E} = (E_1, \ldots, E_m)$ drawn independently from $\mathcal{D}$, as well as bits $B = (b_1, \ldots, b_m)$ where each b_i is 1 with independent probability $\mathrm{Tr}(E_i\rho)$, outputs a hypothesis state σ such that

$$\Pr_{E\in\mathcal{D}}\left[\left|\mathrm{Tr}(E\sigma) - \mathrm{Tr}(E\rho)\right| > \gamma\right] \leq \varepsilon$$

with probability at least $1-\delta$ over $\mathcal{E}$ and B.

[10] Anthony and Bartlett [9] proved a generic lower bound on sample complexity in terms of fat-shattering dimension. However, their bound only implies that

$$m = \Omega\left(\frac{1}{\varepsilon}\left(\frac{n/\gamma^2}{\log^2(n/\gamma^2)} + \log\frac{1}{\delta}\right)\right)$$

measurements are necessary. Using an argument more tailored to our problem, we were able to get rid of the $\log^2(n/\gamma^2)$ factor, improving the dependence on $\mathrm{fat}_{\mathcal{C}_n}(\gamma) \sim n/\gamma^2$ to linear.

To prove Theorem 4.1, it will be helpful to introduce a new parameter that we call the *fine-shattering dimension*. This parameter is like the fat-shattering dimension but with additional restrictions.

Definition 4.2 *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $0 < \gamma \leq \frac{1}{2}$ and $\eta \geq 0$ be real numbers. We say a set $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is (γ, η)-fine-shattered by $\mathcal{C}$ if for all $B \subseteq \{1, \ldots, k\}$, there exists a p-concept $F \in \mathcal{C}$ such that for all $i \in \{1, \ldots, k\}$,*

(i) if $i \notin B$ then $\frac{1}{2} - \gamma - \eta \leq F(s_i) \leq \frac{1}{2} - \gamma$, and

(ii) if $i \in B$ then $\frac{1}{2} + \gamma \leq F(s_i) \leq \frac{1}{2} + \gamma + \eta$.

Then the (γ, η)-fine-shattering dimension of $\mathcal{C}$, or $\mathrm{fine}_{\mathcal{C}}(\gamma, \eta)$, is the maximum k such that some $\{s_1, \ldots, s_k\} \subseteq \mathcal{S}$ is (γ, η)-fine-shattered by $\mathcal{C}$.

Clearly $\mathrm{fine}_{\mathcal{C}}(\gamma, \eta) \leq \mathrm{fat}_{\mathcal{C}}(\gamma)$ for all $\mathcal{C}$, η, and γ. The following theorem lower-bounds sample complexity in terms of fine-shattering dimension.

Theorem 4.3 *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\varepsilon, \delta, \gamma, \eta > 0$. Then provided $\mathrm{fine}_{\mathcal{C}}(\gamma, \eta) \geq 2$ and $\varepsilon, \delta < \frac{1}{4}$, there exists a distribution $\mathcal{D}$ over $\mathcal{S}$ for which the following holds.*

(i) Suppose

$$m \leq \max\left\{\frac{\mathrm{fine}_{\mathcal{C}}(\gamma, \eta) - 1}{64\varepsilon}, \frac{1}{4\varepsilon}\ln\frac{1}{2\delta}\right\},$$

and let $F \in \mathcal{C}$. Then there is no learning algorithm that, given samples $X = (x_1, \ldots, x_m)$ drawn independently from $\mathcal{D}$, as well as real numbers $p_1, \ldots, p_m$ such that $|p_i - F(x_i)| \leq \eta$ for all i, outputs a hypothesis H such that

$$\Pr_{x \in \mathcal{D}}\left[|H(x) - F(x)| \geq \gamma\right] \leq \varepsilon$$

with probability at least $1 - \delta$ over X.

(ii) Suppose

$$m \leq \max\left\{\frac{\mathrm{fine}_{\mathcal{C}}(\gamma, \eta) - 1}{A(\gamma + \eta)^2 \varepsilon}, \frac{1}{4\varepsilon}\ln\frac{1}{2\delta}\right\}$$

where A is some universal constant, and let $F \in \mathcal{C}$. Then there is no learning algorithm that, given samples $X = (x_1, \ldots, x_m)$ drawn independently from $\mathcal{D}$, as well as bits $b_1, \ldots, b_m$ such that $\Pr[b_i = 1] = F(x_i)$, outputs a hypothesis H such that

$$\Pr_{x \in \mathcal{D}}\left[|H(x) - F(x)| \geq \gamma\right] \leq \varepsilon$$

with probability at least $1 - \delta$ over X.

Proof. We start with part (i). Let $k = \mathrm{fine}_{\mathcal{C}}(\gamma, \eta)$, let $S = \{s_1, \ldots, s_k\}$ be any set that is (γ, η)-fine-shattered by $\mathcal{C}$, and let $\mathcal{C}^* \subseteq \mathcal{C}$ be a subclass of size 2^k that (γ, η)-fine-shatters S. Then the function F will be chosen uniformly at random from $\mathcal{C}^*$. Also, the distribution $\mathcal{D}$ will choose s_1 with probability $1 - 4\varepsilon$, and otherwise will choose uniformly at random from $\{s_2, \ldots, s_k\}$. Finally, the learning algorithm will be given $p_i = \frac{1}{2} - \gamma$ if $F(x_i) \leq \frac{1}{2} - \gamma$, or $p_i = \frac{1}{2} + \gamma$ if $F(x_i) \geq \frac{1}{2} + \gamma$.

First suppose $m \leq \frac{1}{4\varepsilon}\ln\frac{1}{2\delta}$. Then the m samples $x_1, \ldots, x_m$ will all equal s_1 with probability at least

$$(1 - 4\varepsilon)^{\frac{1}{4\varepsilon}\ln\frac{1}{2\delta}} \geq 2\delta.$$

Conditioned on this happening, the algorithm certainly fails with probability at least $\frac{1}{2}$.

Next suppose $m \leq \frac{k-1}{64\varepsilon}$. Let r be the number of i's such that $x_i \neq s_1$. Then $\mathrm{EX}[r] = 4\varepsilon m$, and hence

$$\Pr\left[r > \frac{k-1}{4}\right] \leq \frac{4\varepsilon m}{(k-1)/4} \leq \frac{1}{4}$$

by Markov's inequality. Furthermore, conditioned on $r \leq \frac{k-1}{4}$, there are at least $\frac{3}{4}(k-1)$ indices i for which the algorithm has "no information" about $F(x_i)$—in other words, cannot predict whether $F(x_i) \leq \frac{1}{2} - \gamma$ or $F(x_i) \geq \frac{1}{2} + \gamma$ better than the outcome of a fair coin toss. Yet to output an H such that

$$\Pr_{x \in \mathcal{D}} \left[|H(x) - F(x)| \geq \gamma \right] \leq \varepsilon < \frac{1}{4},$$

the algorithm needs to guess correctly for at least $\frac{k-1}{2}$ of these i's. Again by Markov's inequality, it can do this with probability at most $\frac{1/2}{3/4} = \frac{2}{3}$. Hence it fails with overall probability at least $\frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4} > \delta$.

Part (ii) can be proved along the same lines as part (i); we merely give a sketch. If $m \leq \frac{1}{4\varepsilon} \ln \frac{1}{2\delta}$, then the algorithm fails with probability at least δ for the same reason as before. Also, as before, the algorithm must be able to guess $F(s_j)$ with probability at least $1 - \delta$ for $\Omega(k-1)$ indices $j \in \{2, \ldots, k\}$. But this requires $m = \Omega\left(\frac{k-1}{(\gamma+\eta)^2 \varepsilon}\right)$ samples $x_1, \ldots, x_m$. For we can think of each $F(s_j)$ as a coin, whose bias (in the best case) is either $\frac{1}{2} - (\gamma + \eta)$ or $\frac{1}{2} + (\gamma + \eta)$. And it is known that, to decide whether a given coin has bias $\frac{1}{2} - (\gamma + \eta)$ or $\frac{1}{2} + (\gamma + \eta)$ with error probability at most $\delta \ll \frac{1}{2}$, one needs to flip the coin $\Omega\left(\frac{1}{(\gamma+\eta)^2}\right)$ times. ■

To finish the proof, the one remaining step is to lower-bound the fine-shattering dimension of n-qubit quantum states. We will do so using the following result of Ambainis et al. [8].

Theorem 4.4 (Ambainis et al. [8]) *Let $\frac{1}{2} < p < 1$, and let n and k be positive integers satisfying $n \geq (1 - H(p))\, k + 7 \log_2 k$. Then there exist n-qubit mixed states $\{\rho_y\}_{y \in \{0,1\}^k}$ and measurements $E_1, \ldots, E_k$ such that for all $y \in \{0,1\}^k$ and $i \in \{1, \ldots, k\}$:*

(i) if $y_i = 0$ then $\mathrm{Tr}(E_i \rho_y) \leq 1 - p$, and

(ii) if $y_i = 1$ then $\mathrm{Tr}(E_i \rho_y) \geq p$.

Informally, Theorem 4.4 says that Theorem 2.5—the lower bound on quantum encodings due to Nayak [25]—is essentially tight. Incidentally, the encoding scheme of Theorem 4.4 is completely classical, in the sense that the ρ_y's and E_i's are both diagonal in the computational basis. However, we find it more convenient to state the result in quantum language.

We will actually need a slight extension of Theorem 4.4, which bounds the $\mathrm{Tr}(E_i \rho_y)$'s on *both* sides rather than only one.

Theorem 4.5 *Let $\frac{1}{2} < p < 1$, let $\eta > 0$, and let n and k be positive integers satisfying $k \leq \frac{2}{\eta}$ and $n \geq (1 - H(p))\, k + 7 \log_2 \frac{1}{\eta}$. Then there exist n-qubit mixed states $\{\rho_y\}_{y \in \{0,1\}^k}$, and measurements $E_1, \ldots, E_k$, such that for all $y \in \{0,1\}^k$ and $i \in \{1, \ldots, k\}$:*

(i) if $y_i = 0$ then $1 - p - \eta \leq \mathrm{Tr}(E_i \rho_y) \leq 1 - p$, and

(ii) if $y_i = 1$ then $p \leq \mathrm{Tr}(E_i \rho_y) \leq p + \eta$.

Proof Sketch. Follows by small tweaks to the proof of Theorem 4.4 found in [8]. Without going into too much detail, in Ambainis et al.'s construction we first choose a random set $\mathcal{T} = \{(\pi_1, r_1), \ldots, (\pi_\ell, r_\ell)\}$ of transformations of a certain covering code, where $\ell = k^3$. We then use a Chernoff bound to argue that, with high probability over the choice of $\mathcal{T}$, each of the $k 2^k$ probabilities $\mathrm{Tr}(E_i \rho_y)$ satisfies $|\mathrm{Tr}(E_i \rho_y) - q| \leq 1/k$ for some universal constant $q \geq p + 1/k$. It follows that there *exists* a $\mathcal{T}$ for which this property holds.

Now observe that without loss of generality we can make $q = p + 1/k$. To do so, we simply "dilute" each ρ_y to $c\rho_y + (1-c) I$, where I is the maximally mixed state and $c = \frac{p+1/k-1/2}{q-1/2}$. This already proves the theorem in the special case $\eta = 2/k$. If $\eta < 2/k$, then it suffices to repeat Ambainis et al.'s argument with $\ell = 4k/\eta^2$ instead of $\ell = k^3$. ■

As in Section 2, let $\mathcal{S}$ be the set of two-outcome measurements on n qubits, and let $\mathcal{C}_n$ be class of all functions $F : \mathcal{S} \to [0,1]$ such that $F(E) = \mathrm{Tr}(E\rho)$ for some n-qubit mixed state ρ. Then Theorem 4.5 has the following corollary.

Corollary 4.6 *For all positive integers n and all $\gamma \geq \sqrt{n2^{-(n-5)/35}/8}$,*

$$\mathrm{fine}_{\mathcal{C}_n}\left(\gamma, \frac{8\gamma^2}{n}\right) \geq \left\lfloor \frac{n}{5\gamma^2} \right\rfloor.$$

Proof. It is clear that any lower bound on k that we can obtain from Theorem 4.5 is also a lower bound on $\mathrm{fine}_{\mathcal{C}_n}(\gamma, \eta)$. Let $p := \gamma + \frac{1}{2}$. Then basic properties of the entropy function imply that $1 - H(p) \leq 4\gamma^2$. Also, let $k := \lfloor n/5\gamma^2 \rfloor$ and $\eta := 8\gamma^2/n$. Then one can check that the two conditions of Theorem 4.5 are satisfied, as follows. Firstly, $k \leq \frac{n}{4\gamma^2} = \frac{2}{\eta}$. Secondly,

$$\begin{aligned} n &\geq 5\gamma^2 k \\ &\geq (1 - H(p))\, k + \gamma^2 k \\ &\geq (1 - H(p))\, k + \frac{n}{5} - \gamma^2 \\ &\geq (1 - H(p))\, k + \frac{n-5}{5} \\ &\geq (1 - H(p))\, k + 7 \log_2 \frac{1}{\eta}, \end{aligned}$$

where the last line uses the fact that $\eta \geq 2^{-(n-5)/35}$. Hence there exist n-qubit mixed states $\{\rho_y\}_{y \in \{0,1\}^k}$ and measurements $E_1, \ldots, E_k$ such that for all $y \in \{0,1\}^k$ and $i \in \{1, \ldots, k\}$:

(i) if $y_i = 0$ then $\frac{1}{2} - \gamma - \eta \leq \mathrm{Tr}(E_i \rho_y) \leq \frac{1}{2} - \gamma$, and

(ii) if $y_i = 1$ then $\frac{1}{2} + \gamma \leq \mathrm{Tr}(E_i \rho_y) \leq \frac{1}{2} + \gamma + \eta$.

■

Theorem 4.1 now follows immediately by combining Theorem 4.3 with Corollary 4.6.

5 Application to Quantum Communication

In this section we use the quantum learning theorem to prove a new result about *one-way communication complexity*. In this subject we consider two players, Alice and Bob, who hold inputs x and y respectively. For concreteness, let x be an N-bit string, and let y be an M-bit string. Also, let $f : \mathcal{Z} \to \{0,1\}$ be a Boolean function, where $\mathcal{Z}$ is some subset of $\{0,1\}^N \times \{0,1\}^M$. We call f *total* if $\mathcal{Z} = \{0,1\}^N \times \{0,1\}^M$, and *partial* otherwise.

We are interested in the minimum number of bits k that Alice needs to send to Bob, for Bob to be able to evaluate $f(x,y)$ for any input pair $(x,y) \in \mathcal{Z}$. We consider three models of communication: deterministic, randomized, and quantum. In the deterministic model, Alice sends Bob a k-bit string a_x depending only on x. Then Bob, using only a_x and y, must output $f(x,y)$ with certainty. In the randomized model, Alice sends Bob a k-bit string a drawn from a probability distribution $\mathcal{D}_x$. Then Bob must output $f(x,y)$ with probability at least $\frac{2}{3}$ over $a \in \mathcal{D}_x$.[11] In the quantum model, Alice sends Bob a k-qubit mixed state ρ_x. Then Bob, after measuring ρ_x in a basis depending on y, must output $f(x,y)$ with probability at least $\frac{2}{3}$. We use $\mathrm{D}^1(f)$, $\mathrm{R}^1(f)$, and $\mathrm{Q}^1(f)$ to denote the minimum value of k for which Bob can succeed in the deterministic, randomized, and quantum models respectively. Clearly $\mathrm{D}^1(f) \geq \mathrm{R}^1(f) \geq \mathrm{Q}^1(f)$ for all f.

The question that interests us is how small the quantum communication complexity $\mathrm{Q}^1(f)$ can be compared to the classical complexities $\mathrm{D}^1(f)$ and $\mathrm{R}^1(f)$. We know that there exists a total function $f : \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ for which $\mathrm{D}^1(f) = N$ but $\mathrm{R}^1(f) = \mathrm{Q}^1(f) = O(\log N)$.[12] Furthermore, Kerenidis and Raz [22] have recently shown that there exists a partial function f for which $\mathrm{R}^1(f) = \Omega\left(\sqrt{N}\right)$ but $\mathrm{Q}^1(f) = O(\log N)$. A similar result was obtained independently by Gavinsky, Kempe, and de Wolf [17].

[11] We can assume without loss of generality that Bob is deterministic, i.e. that his output is a function of a and y.

[12] This f is the equality function: $f(x,y) = 1$ if $x = y$, and $f(x,y) = 0$ otherwise.

On the other hand, it follows from a result of Klauck [23] that $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ for all *total* f. Intuitively, if Bob's input is small, then quantum communication provides at most a limited savings over classical communication. But does the $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ bound hold for *partial* f as well? Aaronson [2] proved a slightly weaker result: for all f (partial or total), $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\log \mathrm{Q}^1(f)\right)$. Whether the $\log \mathrm{Q}^1(f)$ factor can be removed has remained open for several years. Using our quantum learning theorem, we are able to resolve this question, at the cost of replacing $\mathrm{D}^1(f)$ by $\mathrm{R}^1(f)$. We now prove Theorem 1.4, that $\mathrm{R}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ for any Boolean function f.

Proof of Theorem 1.4. Let $f : \mathcal{Z} \rightarrow \{0,1\}$ be a Boolean function with $\mathcal{Z} \subseteq \{0,1\}^N \times \{0,1\}^M$. Fix Alice's input $x \in \{0,1\}^N$, and let $\mathcal{Z}_x$ be the set of all $y \in \{0,1\}^M$ such that $(x,y) \in \mathcal{Z}$. By Yao's minimax principle, to give a randomized protocol that errs with probability at most $\frac{1}{3}$ for all $y \in \mathcal{Z}_x$, it is enough, for any fixed probability distribution $\mathcal{D}$ over $\mathcal{Z}_x$, to give a randomized protocol that errs with probability at most $\frac{1}{3}$ over y drawn from $\mathcal{D}$.[13]

So let $\mathcal{D}$ be such a distribution; then the randomized protocol is as follows. First Alice chooses k inputs $y_1, \ldots, y_k$ independently from $\mathcal{D}$, where $k = O\left(\mathrm{Q}^1(f)\right)$. She then sends Bob $y_1, \ldots, y_k$, together with $f(x, y_i)$ for all $i \in \{1, \ldots, k\}$. Clearly this message requires only $O\left(M\,\mathrm{Q}^1(f)\right)$ classical bits. We need to show that it lets Bob evaluate $f(x,y)$, with high probability over y drawn from $\mathcal{D}$.

By amplification, we can assume Bob errs with probability at most η for any fixed constant $\eta > 0$. We will take $\eta = \frac{1}{100}$. Also, in the quantum protocol for f, let ρ_x be the $\mathrm{Q}^1(f)$-qubit mixed state that Alice would send given input x, and let E_y be the measurement that Bob would apply given input y. Then $\mathrm{Tr}(E_y \rho_x) \geq 1 - \eta$ if $f(x,y) = 1$, while $\mathrm{Tr}(E_y \rho_x) \leq \eta$ if $f(x,y) = 0$.

Given Alice's classical message, first Bob finds a $\mathrm{Q}^1(f)$-qubit state σ such that $\left|\mathrm{Tr}(E_{y_i}\sigma) - f(x, y_i)\right| \leq \eta$ for all $i \in \{1, \ldots, k\}$. Certainly such a state exists (for take $\sigma = \rho_x$), and Bob can find it by searching exhaustively for its classical description. If there are multiple such states, then Bob chooses one in some arbitrary deterministic way (for example, by lexicographic ordering). Note that we then have $\left|\mathrm{Tr}(E_{y_i}\sigma) - \mathrm{Tr}(E_{y_i}\rho_x)\right| \leq \eta$ for all $i \in \{1, \ldots, k\}$ as well. Finally Bob outputs $f(x,y) = 1$ if $\mathrm{Tr}(E_y\sigma) \geq \frac{1}{2}$, or $f(x,y) = 0$ if $\mathrm{Tr}(E_y\sigma) < \frac{1}{2}$.

Set $\varepsilon = \delta = \frac{1}{6}$ and $\gamma = 0.42$, so that $\gamma\varepsilon = 7\eta$. Then by Theorem 1.1,

$$\Pr_{y \in \mathcal{D}}\left[\left|\mathrm{Tr}(E_y\sigma) - \mathrm{Tr}(E_y\rho_x)\right| > \gamma\right] > \varepsilon$$

with probability at most δ over Alice's classical message, provided that

$$k = \Omega\left(\frac{1}{\gamma^2\varepsilon^2}\left(\frac{\mathrm{Q}^1(f)}{\gamma^2\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\right)\right).$$

So in particular, there exist constants A, B such that if $k \geq A\,\mathrm{Q}^1(f) + B$, then

$$\Pr_{y \in \mathcal{D}}\left[\left|\mathrm{Tr}(E_y\sigma) - f(x,y)\right| > \gamma + \eta\right] > \varepsilon$$

with probability most δ. Since $\gamma + \eta < \frac{1}{2}$, it follows that Bob's classical strategy will fail with probability at most $\varepsilon + \delta = \frac{1}{3}$ over y drawn from $\mathcal{D}$. ■

It is easy to see that, in Theorem 1.4, the upper bound on $\mathrm{R}^1(f)$ needs to depend both on M and on $\mathrm{Q}^1(f)$. For the index function[14] yields a total f for which $\mathrm{R}^1(f)$ is exponentially larger than M, while the recent results of Kerenidis and Raz [22] and of Gavinsky et al. [17] yield a partial f for which $\mathrm{R}^1(f)$ is exponentially larger than $\mathrm{Q}^1(f)$. However, is it possible that Theorem 1.4 could be improved to $\mathrm{R}^1(f) = O\left(M + \mathrm{Q}^1(f)\right)$?

Using a slight generalization of Gavinsky et al.'s result, we are able to rule out this possibility. Gavinsky et al. consider the following one-way communication problem, called the *Boolean Hidden Matching Problem.* Alice is given a string $x \in \{0,1\}^N$. For some parameter $\alpha > 0$, Bob is given αN disjoint edges $(i_1, j_1), \ldots, (i_{\alpha N}, j_{\alpha N})$ in $\{1, \ldots, N\}^2$, together with a string $w \in \{0,1\}^{\alpha N}$. (Thus Bob's input length is $M = O(\alpha N \log N)$.) Alice and Bob are promised that either

[13] Indeed, it suffices to give a *deterministic* protocol that errs with probability at most $\frac{1}{3}$ over y drawn from $\mathcal{D}$, a fact we will not need.

[14] This is the function $f : \{0,1\}^N \times \{1, \ldots, N\} \rightarrow \{0,1\}$ defined by $f(x_1 \cdots x_N, i) = x_i$.

(i) $x_{i_\ell} \oplus x_{j_\ell} \equiv w_\ell \pmod 2$ for all $\ell \in \{1,\ldots,\alpha N\}$, or

(ii) $x_{i_\ell} \oplus x_{j_\ell} \not\equiv w_\ell \pmod 2$ for all $\ell \in \{1,\ldots,\alpha N\}$.

Bob's goal is to output $f = 0$ in case (i), or $f = 1$ in case (ii).

It is not hard to see that $\mathrm{Q}^1(f) = O\left(\frac{1}{\alpha}\log N\right)$ for all $\alpha > 0$.[15] What Gavinsky et al. showed is that, if $\alpha \approx 1/\sqrt{\log N}$, then $\mathrm{R}^1(f) = \Omega\left(\sqrt{N/\alpha}\right)$. By tweaking their proof a bit, one can generalize their result to $\mathrm{R}^1(f) = \Omega\left(\sqrt{N/\alpha}\right)$ for *all* $\alpha \ll 1/\sqrt{\log N}$.[16] So in particular, set $\alpha := 1/\sqrt{N}$. Then we obtain a partial Boolean function f for which $M = O\left(\sqrt{N}\log N\right)$ and $\mathrm{Q}^1(f) = O\left(\sqrt{N}\log N\right)$ but $\mathrm{R}^1(f) = \Omega\left(N^{3/4}\right)$, thereby refuting the conjecture that $\mathrm{R}^1(f) = O\left(M + \mathrm{Q}^1(f)\right)$.

As a final remark, the Boolean Hidden Matching Problem clearly satisfies $\mathrm{D}^1(f) = \Omega(N)$ for all $\alpha > 0$. So by varying α, we immediately get not only that $\mathrm{D}^1(f) = O\left(M + \mathrm{Q}^1(f)\right)$ is false, but that Aaronson's bound $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\log \mathrm{Q}^1(f)\right)$ [2] is *tight* up to a polylogarithmic term. This answers one of the open questions in [2].

6 Application to Quantum Advice

Having applied our quantum learning theorem to communication complexity, in this section we apply the theorem to computational complexity. In particular, we will show how to use a trusted classical string to perform approximate verification of an untrusted quantum state.

The following conventions will be helpful throughout the section. We identify a language $L \subseteq \{0,1\}^*$ with the Boolean function $L : \{0,1\}^* \rightarrow \{0,1\}$ such that $L(x) = 1$ if and only if $x \in L$. Given a quantum algorithm A, we let $P_A^1(|\psi\rangle)$ be the probability that A accepts and $P_A^0(|\psi\rangle)$ be the probability that A rejects if given the state $|\psi\rangle$ as input. Note that A might neither accept nor reject (in other words, output "don't know"), in which case $P_A^0(|\psi\rangle) + P_A^1(|\psi\rangle) < 1$. Finally, we use $\mathcal{H}_2^{\otimes k}$ to denote a Hilbert space of k qubits, and $\mathrm{poly}(n)$ to denote an arbitrary polynomial in n.

6.1 Quantum Advice and Proofs

Recall that BQP, or Bounded-Error Quantum Polynomial-Time, is the class of problems efficiently solvable by a quantum computer. Then BQP/qpoly is a generalization of BQP, in which the quantum computer is given a polynomial-size "quantum advice state" that depends only on the input length n, but could otherwise be arbitrarily hard to prepare. More formally:

Definition 6.1 *A language $L \subseteq \{0,1\}^*$ is in* BQP/qpoly *if there exists a polynomial-time quantum algorithm A such that for all input lengths n, there exists a quantum advice state $|\psi_n\rangle \in \mathcal{H}_2^{\otimes \mathrm{poly}(n)}$ such that $P_A^{L(x)}(|x\rangle|\psi_n\rangle) \geq \frac{2}{3}$ for all $x \in \{0,1\}^n$.*

How powerful is this class? Aaronson [2] proved the first limitation on BQP/qpoly, by showing that BQP/qpoly $\subseteq$ PostBQP/poly. Here PostBQP is a generalization of BQP in which we can "postselect" on the outcomes of measurements,[17] and /poly means "with polynomial-size classical advice." Intuitively, this result means that anything we can do with quantum advice, we can also do with *classical* advice, provided we are willing to use exponentially more computation time to extract what the advice is telling us.

In addition to quantum advice, we will also be interested in quantum proofs. Compared to advice, a proof has the advantage that it can be tailored to a particular input x, but the disadvantage that it cannot be trusted. In other words, while an advisor's only goal is to help the algorithm A decide whether $x \in L$, a prover wants to *convince* A that $x \in L$. The class of problems that admit polynomial-size quantum proofs is called QMA (Quantum Merlin-Arthur).

[15] The protocol is as follows: first Alice sends the $\log N$-qubit quantum message $\frac{1}{\sqrt{N}}\sum_{i=1}^N (-1)^{x_i}|i\rangle$. Then Bob measures in a basis corresponding to $(i_1,j_1),\ldots,(i_{\alpha N},j_{\alpha N})$. With probability 2α, Bob will learn whether $x_{i_\ell} \oplus x_{j_\ell} \equiv w_\ell$ for some edge (i_ℓ,j_ℓ). So it suffices to amplify the protocol $O(1/\alpha)$ times.

[16] R. de Wolf, personal communication.

[17] See [3] for a detailed definition, as well as a proof that PostBQP coincides with the classical complexity class PP.

Definition 6.2 *A language L is in* QMA *if there exists a polynomial-time quantum algorithm A such that for all $x \in \{0,1\}^n$:*

(i) If $x \in L$ then there exists a quantum witness $|\varphi\rangle \in \mathcal{H}_2^{\otimes \text{poly}(n)}$ such that $P_A^1(|x\rangle|\varphi\rangle) \geq \frac{2}{3}$.

(ii) If $x \notin L$ then $P_A^1(|x\rangle|\varphi\rangle) \leq \frac{1}{3}$ for all $|\varphi\rangle$.

One can think of QMA as a quantum analogue of NP.

6.2 Untrusted Advice

To state our result in the strongest possible way, we need to define a new notion called *untrusted advice*, which might be of independent interest for complexity theory. Intuitively, untrusted advice is a "hybrid" of proof and advice: it is like a proof in that it cannot be trusted, but like advice in that depends only on the input length n. More concretely, let us define the complexity class YP, or "Yoda Polynomial-Time," to consist of all problems solvable in classical polynomial time with help from polynomial-size untrusted advice:[18]

Definition 6.3 *A language L is in* YP *if there exists a polynomial-time algorithm A such that for all n:*

(i) There exists a string $y_n \in \{0,1\}^{p(n)}$ such that $A(x, y_n)$ outputs $L(x)$ for all $x \in \{0,1\}^n$.

(ii) $A(x,y)$ outputs either $L(x)$ or "don't know" for all $x \in \{0,1\}^n$ and all y.

From the definition, it is clear that YP is contained both in P/poly and in NP $\cap$ coNP. In Appendix 10, we prove several other results relating YP to standard complexity classes.

Naturally one can also define YPP and YQP, the (bounded-error) probabilistic and quantum analogues of YP. For brevity, we give only the definition of YQP.

Definition 6.4 *A language L is in* YQP *if there exists a polynomial-time quantum algorithm A such that for all n:*

(i) There exists a state $|\varphi_n\rangle \in \mathcal{H}_2^{\otimes \text{poly}(n)}$ such that $P_A^{L(x)}(|x\rangle|\varphi_n\rangle) \geq \frac{2}{3}$ for all $x \in \{0,1\}^n$.

(ii) $P_A^{1-L(x)}(|x\rangle|\varphi\rangle) \leq \frac{1}{3}$ for all $x \in \{0,1\}^n$ and all $|\varphi\rangle$.

By analogy to the classical case, YQP is contained both in BQP/qpoly and in QMA $\cap$ coQMA. We also have YQP/qpoly = BQP/qpoly, since the untrusted YQP advice can be tacked onto the trusted /qpoly advice. Figure 1 shows the known containments among various classes involving quantum advice and proofs.

6.3 Heuristic Complexity

Ideally, we would like to show that BQP/qpoly = YQP/poly—in other words, that trusted quantum advice can be replaced by trusted classical advice together with untrusted quantum advice. However, we will only be able to prove this for the *heuristic* versions of these classes: that is, the versions where we allow algorithms that can err on some fraction of inputs.[19] We now explain what this means (for details, see the excellent survey by Bogdanov and Trevisan [13]).

A *distributional problem* is a pair $(L, \{\mathcal{D}_n\})$, where $L \subseteq \{0,1\}^*$ is a language and $\mathcal{D}_n$ is a probability distribution over $\{0,1\}^n$. Intuitively, for each input length n, the goal will be to decide whether $x \in L$ with high probability over x drawn from $\mathcal{D}_n$. In particular, the class HeurP, or Heuristic-P, consists (roughly speaking) of all distributional problems that can be solved in polynomial time on a $1 - \frac{1}{\text{poly}(n)}$ fraction of inputs.

[18] Here Yoda, from *Star Wars*, is intended to evoke a sage whose messages are highly generic ("Do or do not... there is no try"). One motivation for the name YP is that, to our knowledge, there had previously been no complexity class starting with a 'Y'.

[19] Closely related to heuristic complexity is the better-known *average-case* complexity. In average-case complexity one considers algorithms that can never err, but that are allowed to output "don't know" on some fraction of inputs.

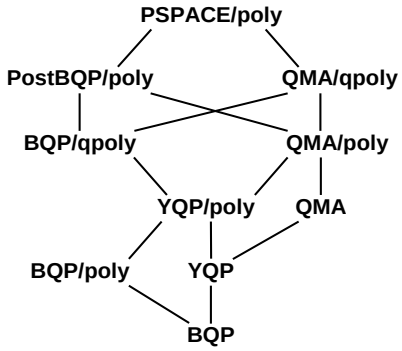

Figure 1: Some quantum advice and proof classes. The containment BQP/qpoly $\subseteq$ PostBQP/poly was shown in [2], while QMA/qpoly $\subseteq$ PSPACE/poly was shown in [4].

Definition 6.5 *A distributional problem $(L, \{\mathcal{D}_n\})$ is in* HeurP *if there exists a polynomial-time algorithm A such that for all n and $\varepsilon > 0$:*

$$\Pr_{x \in \mathcal{D}_n} \left[A\left(x, 0^{\lceil 1/\varepsilon \rceil}\right) \text{ outputs } L(x) \right] \geq 1 - \varepsilon.$$

One can also define HeurP/poly, or HeurP with polynomial-size advice. (Note that in this context, "polynomial-size" means polynomial not just in n but in $1/\varepsilon$ as well.) Finally, let us define the heuristic analogues of BQP and YQP.

Definition 6.6 *A distributional problem $(L, \{\mathcal{D}_n\})$ is in* HeurBQP *if there exists a polynomial-time quantum algorithm A such that for all n and $\varepsilon > 0$:*

$$\Pr_{x \in \mathcal{D}_n} \left[P_A^{L(x)} \left(|x\rangle |0\rangle^{\otimes \lceil 1/\varepsilon \rceil} \right) \geq \frac{2}{3} \right] \geq 1 - \varepsilon.$$

Definition 6.7 *A distributional problem $(L, \{\mathcal{D}_n\})$ is in* HeurYQP *if there exists a polynomial-time quantum algorithm A such that for all n and $\varepsilon > 0$:*

(i) There exists a state $|\varphi_{n,\varepsilon}\rangle \in \mathcal{H}_2^{\otimes \operatorname{poly}(n, 1/\varepsilon)}$ such that

$$\Pr_{x \in \mathcal{D}_n} \left[P_A^{L(x)} \left(|x\rangle |\varphi_{n,\varepsilon}\rangle \right) \geq \frac{2}{3} \right] \geq 1 - \varepsilon.$$

(ii) The probability over $x \in \mathcal{D}_n$ that there exists a $|\varphi\rangle$ such that $P_A^{1-L(x)}(|x\rangle |\varphi\rangle) \geq \frac{1}{3}$ is at most ε.

It is clear that HeurYQP/poly $\subseteq$ HeurBQP/qpoly $=$ HeurYQP/qpoly.

6.4 The Witness Protection Lemma

Our goal is to show that HeurBQP/qpoly $=$ HeurYQP/poly: in the heuristic setting, trusted classical advice can be used to verify untrusted quantum advice. The intuition behind this result is simple: the classical advice to the HeurYQP verifier V will consist of a polynomial number of randomly-chosen "test inputs" $x_1, \ldots, x_m$, as well as whether each x_i belongs to the language L. Then given an untrusted quantum advice state $|\varphi\rangle$, first V will check that $|\varphi\rangle$ yields the correct answers on $x_1, \ldots, x_m$; only if $|\varphi\rangle$ passes this initial test will V use it on the input x of interest. By appealing to our quantum learning theorem, we will argue that any $|\varphi\rangle$ that passes the initial test must yield the correct answers for *most* x with high probability.

But there is a problem: what if a dishonest prover sends a state $|\varphi\rangle$ such that, while V's measurements succeed in "verifying" $|\varphi\rangle$, they also *corrupt* it? Indeed, even if V repeats the verification procedure many times, conceivably $|\varphi\rangle$ could be corrupted by the very last repetition without V ever realizing it. Intuitively, the easiest way to avoid this problem is just to repeat the verification procedure a *random* number of times. To formalize this idea, we will need the following "quantum union bound," which was proved by Aaronson [4] based on a result of Ambainis et al. [8].

Proposition 6.8 (Aaronson [4]) *Let $E_1, \dots, E_m$ be two-outcome measurements, and suppose $\mathrm{Tr}(E_i\rho) \geq 1-\epsilon$ for all $i \in \{1,\dots,m\}$. Then if we apply $E_1, \dots, E_m$ in sequence to the initial state ρ, the probability that any of the E_i's reject is at most $m\sqrt{\epsilon}$.*

Using Proposition 6.8, we can now prove the Witness Protection Lemma.

Lemma 6.9 (Witness Protection Lemma) *Let $\mathcal{E} = \{E_1, \dots, E_m\}$ be a set of two-outcome measurements and let T be a positive integer. Then there exists a test procedure Q with the following properties:*

(i) Q takes a state ρ_0 as input, applies at most T measurements from $\mathcal{E}$, and then returns either "success" or "failure."

(ii) If $\mathrm{Tr}(E_i\rho_0) \geq 1-\epsilon$ for all i, then Q succeeds with probability at least $1-T\sqrt{\epsilon}$.

(iii) If Q succeeds with probability at least λ, then conditioned on succeeding, Q outputs a state σ such that $\mathrm{Tr}(E_i\sigma) \geq 1 - 2\sqrt{\frac{m}{\lambda T}}$ for all i.

Proof. The procedure Q is given by the following pseudocode:

```
Let ρ := ρ_0
Choose t ∈ {1,...,T} uniformly at random
For u := 1 to t
    Choose i ∈ {1,...,m} uniformly at random
    Apply E_i to ρ
    If E_i rejects, return "FAILURE" and halt
Next u
Return "SUCCESS" and output σ := ρ
```

Property (ii) follows immediately from Proposition 6.8. For property (iii), let ρ_u be the state of ρ immediately after the u^{th} iteration, conditioned on iterations $1,\dots,u$ all succeeding. Also, let $\beta_u := \max_i \{1 - \mathrm{Tr}(E_i\rho_u)\}$. Then Q fails in the $(u+1)^{st}$ iteration with probability at least β_u/m, conditioned on succeeding in iterations $1,\dots,u$. So letting p_t be the probability that Q completes all t iterations, we have

$$p_t \leq \left(1 - \frac{\beta_0}{m}\right) \cdots \left(1 - \frac{\beta_{t-1}}{m}\right).$$

Hence, letting $z > 0$ be a parameter to be determined later,

$$\begin{aligned}
\sum_{t\,:\,\beta_t > z} p_t &\leq \sum_{t\,:\,\beta_t > z} \left(1 - \frac{\beta_0}{m}\right) \cdots \left(1 - \frac{\beta_{t-1}}{m}\right) \\
&\leq \sum_{t\,:\,\beta_t > z} \prod_{u < t\,:\,\beta_u > z} \left(1 - \frac{\beta_u}{m}\right) \\
&\leq \sum_{t=0}^{\infty} \left(1 - \frac{z}{m}\right)^t \\
&= \frac{m}{z}.
\end{aligned}$$

Also, by the assumption that Q succeeds with probability at least λ, we have $\frac{1}{T}\sum_t p_t \geq \lambda$. So for all i,

$$\begin{aligned}
1 - \mathrm{Tr}\left(E_i\sigma\right) &= \frac{\sum_t p_t \left(1 - \mathrm{Tr}\left(E_i\rho_t\right)\right)}{\sum_t p_t} \\
&= \frac{\sum_{t \,:\, \beta_t \leq z} p_t \left(1 - \mathrm{Tr}\left(E_i\rho_t\right)\right)}{\sum_t p_t} + \frac{\sum_{t \,:\, \beta_t > z} p_t \left(1 - \mathrm{Tr}\left(E_i\rho_t\right)\right)}{\sum_t p_t} \\
&\leq \frac{\sum_{t \,:\, \beta_t \leq z} p_t\beta_t}{\sum_t p_t} + \frac{m/z}{\sum_t p_t} \\
&\leq z + \frac{m/z}{\lambda T}.
\end{aligned}$$

The last step is to set $z := \sqrt{\frac{m}{\lambda T}}$, thereby obtaining the optimal lower bound

$$\mathrm{Tr}\left(E_i\sigma\right) \geq 1 - 2\sqrt{\frac{m}{\lambda T}}.$$

■

6.5 Main Result

We are now ready to prove Theorem 1.5, that $\mathsf{HeurBQP/qpoly} = \mathsf{HeurYQP/poly}$.

Proof of Theorem 1.5. Fix a distributional problem $(L, \{\mathcal{D}_n\}) \in \mathsf{HeurBQP/qpoly}$. Then there exists a polynomial-time quantum algorithm A such that for all n and $\varepsilon > 0$, there exists a state $|\psi_{n,\varepsilon}\rangle$ of size $q = O\left(\mathrm{poly}\left(n, 1/\varepsilon\right)\right)$ such that

$$\Pr_{x\in\mathcal{D}_n}\left[P_A^{L(x)}\left(|x\rangle\,|\psi_{n,\varepsilon}\rangle\right) \geq \frac{2}{3}\right] \geq 1 - \varepsilon.$$

Let $\mathcal{D}_n^*$ be the distribution obtained by starting from $\mathcal{D}_n$ and then conditioning on $P_A^{L(x)}\left(|x\rangle\,|\psi_{n,\varepsilon}\rangle\right) \geq \frac{2}{3}$. Then our goal will be to construct a polynomial-time verification procedure V such that, for all n and $\varepsilon > 0$, there exists an advice string $a_{n,\varepsilon} \in \{0,1\}^{\mathrm{poly}(n,1/\varepsilon)}$ for which the following holds.

- There exists a state $|\varphi_{n,\varepsilon}\rangle \in \mathcal{H}_2^{\otimes\,\mathrm{poly}(n,1/\varepsilon)}$ such that

$$\Pr_{x\in\mathcal{D}_n^*}\left[P_V^{L(x)}\left(|x\rangle\,|\varphi_{n,\varepsilon}\rangle\,|a_{n,\varepsilon}\rangle\right) \geq \frac{2}{3}\right] \geq 1 - \varepsilon.$$

- The probability over $x \in \mathcal{D}_n^*$ that there exists a state $|\varphi\rangle$ such that $P_V^{1-L(x)}\left(|x\rangle\,|\varphi\rangle\,|a_{n,\varepsilon}\rangle\right) \geq \frac{1}{3}$ is at most ε.

If V succeeds with probability at least $1 - \varepsilon$ over $x \in \mathcal{D}_n^*$, then by the union bound it succeeds with probability at least $1 - 2\varepsilon$ over $x \in \mathcal{D}_n$. Clearly this suffices to prove the theorem.

As a preliminary step, let us replace A by an amplified algorithm A^*, which takes $|\psi_{n,\varepsilon}\rangle^{\otimes\ell}$ as advice and returns the majority answer among ℓ invocations of A. Here ℓ is a parameter to be determined later. By a Chernoff bound,

$$\Pr_{x\in\mathcal{D}_n}\left[P_{A^*}^{L(x)}\left(|x\rangle\,|\psi_{n,\varepsilon}\rangle^{\otimes\ell}\right) \geq 1 - e^{-\ell/18}\right] \geq 1 - \varepsilon.$$

We now describe the verifier V. The verifier receives three objects as input:

- **An input $x \in \{0,1\}^n$.**
- **An untrusted quantum advice state $|\varphi_0\rangle$.** This $|\varphi_0\rangle$ is divided into ℓ registers, each with q qubits. The state that the verifier *expects* to receive is $|\varphi_0\rangle = |\psi_{n,\varepsilon}\rangle^{\otimes\ell}$.

- **A trusted classical advice string $a_{n,\varepsilon}$.** This $a_{n,\varepsilon}$ consists of m test inputs $x_1, \ldots, x_m \in \{0,1\}^n$, together with $L(x_i)$ for $i \in \{1, \ldots, m\}$. Here m is a parameter to be determined later.

Given these objects, V does the following, where T is another parameter to be determined later.

Phase 1: Verify $|\varphi_0\rangle$
Let $|\varphi\rangle := |\varphi_0\rangle$
Choose $t \in \{1, \ldots, T\}$ uniformly at random
For $u := 1$ to t
    Choose $i \in \{1, \ldots, m\}$ uniformly at random
    Simulate $A^*(|x_i\rangle |\varphi\rangle)$
    If A^* outputs $1 - L(x_i)$, output "don't know" and halt
Next u

Phase 2: Decide whether $x \in L$
Simulate $A^*(|x\rangle |\varphi\rangle)$
Accept if A^* outputs 1; reject otherwise

It suffices to show that there exists a choice of test inputs $x_1, \ldots, x_m$, as well as parameters ℓ, m, and T, for which the following holds.

(a) If $|\varphi_0\rangle = |\psi_{n,\varepsilon}\rangle^{\otimes \ell}$, then Phase 1 succeeds with probability at least $\frac{5}{6}$.

(b) If Phase 1 succeeds with probability at least $\frac{1}{3}$, then conditioned on its succeeding, $P_{A^*}^{L(x)}(|x_i\rangle |\varphi\rangle) \geq \frac{17}{18}$ for all $i \in \{1, \ldots, m\}$.

(c) If $P_{A^*}^{L(x)}(|x_i\rangle |\varphi\rangle) \geq \frac{17}{18}$ for all $i \in \{1, \ldots, m\}$, then

$$\Pr_{x \in \mathcal{D}_n^*}\left[P_{A^*}^{L(x)}(|x\rangle |\varphi\rangle) \geq \frac{5}{6}\right] \geq 1 - \varepsilon.$$

For conditions (a)-(c) ensure that the following holds with probability at least $1 - \varepsilon$ over $x \in \mathcal{D}_n^*$. First, if $|\varphi_0\rangle = |\psi_{n,\varepsilon}\rangle^{\otimes \ell}$, then

$$P_V^{L(x)}(|x\rangle |\varphi_0\rangle |a_{n,\varepsilon}\rangle) \geq \frac{5}{6} - \frac{1}{6} = \frac{2}{3}$$

by the union bound. Here $\frac{1}{6}$ is the maximum probability of failure in Phase 1, while $\frac{5}{6}$ is the minimum probability of success in Phase 2. Second, for all $|\varphi_0\rangle$, either Phase 1 succeeds with probability less than $\frac{1}{3}$, or else Phase 2 succeeds with probability at least $\frac{5}{6}$. Hence

$$P_V^{1-L(x)}(|x\rangle |\varphi_0\rangle |a_{n,\varepsilon}\rangle) \leq \max\left\{\frac{1}{3}, \frac{1}{6}\right\} = \frac{1}{3}.$$

Therefore V is a valid HeurYQP/poly verifier as desired.

Set

$$\begin{aligned} m &:= K\frac{q}{\varepsilon}\log^3\frac{q}{\varepsilon}, \\ \ell &:= 100 + 9\ln m, \\ T &:= 3888m, \end{aligned}$$

where $K > 0$ is a sufficiently large constant and q is the number of qubits of $|\psi_{n,\varepsilon}\rangle$. Also, form the advice string $a_{n,\varepsilon}$ by choosing $x_1, \ldots, x_m$ independently from $\mathcal{D}_n^*$. We will show that conditions (a)-(c) all hold with high probability over the choice of $x_1, \ldots, x_m$—and hence, that there certainly *exists* a choice of $x_1, \ldots, x_m$ for which they hold.

To prove (a), we appeal to part (ii) of Lemma 6.9. Setting $\epsilon := e^{-\ell/18}$, we have $P_{A^*}^{L(x)}\left(|x_i\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right) \geq 1-\epsilon$ for all $i \in \{1,\ldots,m\}$. Therefore Phase 1 succeeds with probability at least

$$1 - T\sqrt{\epsilon} = 1 - 3888m \cdot e^{-\ell/9} \geq \frac{5}{6}.$$

To prove (b), we appeal to part (iii) of Lemma 6.9. Set $\lambda := \frac{1}{3}$. Then if Phase 1 succeeds with probability at least λ, for all i we have

$$P_{A^*}^{L(x)}\left(|x_i\rangle |\varphi\rangle\right) \geq 1 - 2\sqrt{\frac{m}{\lambda T}} = 1 - 2\sqrt{\frac{3m}{3888m}} = \frac{17}{18}.$$

Finally, to prove (c), we appeal to Theorem 1.2. Set $\eta := \frac{1}{18}$. Then for all i we have

$$P_{A^*}^{L(x)}\left(|x_i\rangle |\varphi\rangle\right) \geq \frac{17}{18} = 1 - \eta,$$

and also

$$P_{A^*}^{L(x)}\left(|x_i\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right) \geq 1 - e^{-\ell/18} > 1 - \eta.$$

Hence

$$\left|P_{A^*}^{L(x)}\left(|x_i\rangle |\varphi\rangle\right) - P_{A^*}^{L(x)}\left(|x_i\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right)\right| \leq \eta.$$

Now set $\gamma := \frac{1}{9}$ and $\delta := \frac{1}{3}$. Then $\gamma > \eta$ and

$$\begin{aligned} m &= \Omega\left(\frac{q}{\varepsilon}\log^3\frac{q}{\varepsilon}\right) \\ &= \Omega\left(\frac{q\ell}{\varepsilon}\log^2\frac{q\ell}{\varepsilon}\right) \\ &= \Omega\left(\frac{1}{\varepsilon}\left(\frac{q\ell}{(\gamma-\eta)^2}\log^2\frac{q\ell}{(\gamma-\eta)\,\varepsilon} + \log\frac{1}{\delta}\right)\right). \end{aligned}$$

So Theorem 1.2 implies that

$$\Pr_{x\in\mathcal{D}_n^*}\left[\left|P_{A^*}^{L(x)}\left(|x\rangle |\varphi\rangle\right) - P_{A^*}^{L(x)}\left(|x\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right)\right| > \gamma\right] \leq \varepsilon$$

and hence

$$\Pr_{x\in\mathcal{D}_n^*}\left[P_{A^*}^{L(x)}\left(|x\rangle |\varphi\rangle\right) < \frac{5}{6}\right] \leq \varepsilon$$

with probability at least $1-\delta$ over the choice of $a_{n,\varepsilon}$. Here we have used the facts that

$$P_{A^*}^{L(x)}\left(|x\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right) \geq 1 - \eta$$

and that $\eta + \gamma = \frac{1}{18} + \frac{1}{9} = \frac{1}{6}$. ■

7 Open Problems

Perhaps the central question left open by this paper is which classes of states and measurements can be learned, not only with a linear number of measurements, but also with a reasonable amount of computation. To give two examples, what is the situation for stabilizer states [5] or noninteracting-fermion states [31]?[20]

[20]Note that we can only hope to learn such states efficiently for restricted classes of measurements. Otherwise, even if the state to be learned were a classical basis state $|x\rangle$, a "measurement" of $|x\rangle$ might be an arbitrary polynomial-time computation that fed x as input to a pseudorandom function.

On the experimental side, it would be interesting to demonstrate "pretty good tomography" in photonics, ion traps, NMR, or any other technology that allows the preparation and measurement of multi-qubit entangled states. Already for 3 or 4 qubits, complete tomography requires hundreds of measurements, and depending on what accuracy is needed, it seems likely that our learning approach could yield an efficiency improvement.

It would also be useful to generalize our results—most obviously, to k-outcome measurements and noisy measurements. One might hope for an even broader generalization, to what is known as *quantum process tomography*. Here the goal is to learn an unknown quantum *operation* on n qubits by feeding it inputs and examining the outputs. But for process tomography, it is not hard to show that exponentially many measurements really *are* needed; in other words, the analogue of our learning theorem is false.[21] Still, it would be interesting to know if there is anything to say about "pretty good process tomography" for restricted classes of operations.

Finally, our quantum information results immediately suggest several problems. First, does $\mathsf{BQP/qpoly} = \mathsf{YQP/poly}$? In other words, can we use classical advice to verify quantum advice even in the worst-case setting? Alternatively, can we give a "quantum oracle" (see [6]) relative to which $\mathsf{BQP/qpoly} \neq \mathsf{YQP/poly}$? Second, can the relation $\mathrm{R}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ be improved to $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ for all f? Perhaps learning theory techniques could even shed light on the old problem of whether $\mathrm{R}^1(f) = O\left(\mathrm{Q}^1(f)\right)$ for all total f.

8 Acknowledgments

I thank Noga Alon, Peter Bartlett, Aram Harrow, Tony Leggett, Peter Shor, Aephraim Steinberg, Luca Trevisan, and Ronald de Wolf for helpful discussions and correspondence.

References

[1] S. Aaronson. Multilinear formulas and skepticism of quantum computing. In *Proc. ACM STOC*, pages 118–127, 2004. Journal version to appear in SICOMP. quant-ph/0311039.

[2] S. Aaronson. Limitations of quantum advice and one-way communication. *Theory of Computing*, 1:1–28, 2005. quant-ph/0402095.

[3] S. Aaronson. Quantum computing, postselection, and probabilistic polynomial-time. *Proc. Roy. Soc. London*, A461(2063):3473–3482, 2005. quant-ph/0412187.

[4] S. Aaronson. QMA/qpoly is contained in PSPACE/poly: de-Merlinizing quantum protocols. In *Proc. IEEE Conference on Computational Complexity*, pages 261–273, 2006. quant-ph/0510230.

[5] S. Aaronson and D. Gottesman. Improved simulation of stabilizer circuits. *Phys. Rev. Lett.*, 70(052328), 2004. quant-ph/0406196.

[6] S. Aaronson and G. Kuperberg. Quantum versus classical proofs and advice. Submitted. quant-ph/0604056, 2006.

[7] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

[8] A. Ambainis, A. Nayak, A. Ta-Shma, and U. V. Vazirani. Quantum dense coding and quantum finite automata. *J. ACM*, 49:496–511, 2002. Earlier version in ACM STOC 1999, pp. 376-383. quant-ph/9804043.

[9] M. Anthony and P. Bartlett. Function learning from interpolation. *Combinatorics, Probability, and Computing*, 9(3):213–225, 2000.

[21] Here is a proof sketch: let U be an n-qubit unitary that maps $|x\rangle|b\rangle$ to $|x\rangle|b \oplus f(x)\rangle$, for some Boolean function $f : \{0,1\}^{n-1} \rightarrow \{0,1\}$. Then to predict U on a $1-\varepsilon$ fraction of basis states, we need to know $(1-\varepsilon)\,2^{n-1}$ bits of the truth table of f. But Holevo's Theorem implies that, by examining $U|\psi_i\rangle$ for T input states $|\psi_1\rangle, \ldots, |\psi_T\rangle$, we can learn at most Tn bits about f.

[10] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *J. Comput. Sys. Sci.*, 56(2):174–190, 1998.

[11] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *J. Comput. Sys. Sci.*, 52(3):434–452, 1996.

[12] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[13] A. Bogdanov and L. Trevisan. Average-case complexity. ECCC TR06-073, 2006.

[14] V. Bužek. Quantum tomography from incomplete data via MaxEnt principle. In [28]. Pages 189-234.

[15] V. Bužek, G. Drobný, R. Derka, G. Adam, and H. Wiedemann. Quantum state reconstruction from incomplete data. *Chaos, Solitons, and Fractals*, 10(6):981–1074, 1999. quant-ph/9805020.

[16] G. D'Ariano, M. De Laurentis, M. Paris, A. Porzio, and S. Solimeno. Quantum tomography as a tool for the characterization of optical devices. *Journal of Optics B: Quantum and Semiclassical Optics*, 4:S127–S132, 2002. quant-ph/0110110.

[17] D. Gavinsky, J. Kempe, and R. de Wolf. Exponential separation of quantum and classical one-way communication complexity for a Boolean function. ECCC TR06-086, quant-ph/0607174, 2006.

[18] O. Goldreich. On quantum computing. www.wisdom.weizmann.ac.il/~oded/on-qc.html, 2004.

[19] O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions. *J. ACM*, 33(4):792–807, 1984.

[20] A. S. Holevo. Some estimates of the information transmitted by quantum communication channels. *Problems of Information Transmission*, 9:177–183, 1973. English translation.

[21] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Sys. Sci.*, 48(3):464–497, 1994.

[22] I. Kerenidis and R. Raz. The one-way communication complexity of the Boolean Hidden Matching Problem. ECCC TR06-087, quant-ph/0607173, 2006.

[23] H. Klauck. Quantum communication complexity. In *Proc. Intl. Colloquium on Automata, Languages, and Programming (ICALP)*, pages 241–252, 2000. quant-ph/0005032.

[24] L. A. Levin. Polynomial time and extravagant models, in The tale of one-way functions. *Problems of Information Transmission*, 39(1):92–103, 2003. cs.CR/0012023.

[25] A. Nayak. Optimal lower bounds for quantum automata and random access codes. In *Proc. IEEE FOCS*, pages 369–377, 1999. quant-ph/9904093.

[26] H. Nishimura and T. Yamakami. Polynomial time quantum computation with advice. *Inform. Proc. Lett.*, 90:195–204, 2003. ECCC TR03-059, quant-ph/0305100.

[27] J. L. O'Brien, G. J. Pryde, A. G. White, T. C. Ralph, and D. Branning. Demonstration of an all-optical quantum controlled-NOT gate. *Nature*, 426:264–267, 2003. quant-ph/0403062.

[28] M. G. A. Paris and J. Řeháček, editors. *Quantum State Estimation*. Springer, 2004.

[29] K. Resch, P. Walther, and A. Zeilinger. Full characterization of a three-photon Greenberger-Horne-Zeilinger state using quantum state tomography. *Phys. Rev. Lett.*, 94(070402), 2005. quant-ph/0412151.

[30] E. Skovsen, H. Stapelfeldt, S. Juhl, and K. Mølmer. Quantum state tomography of dissociating molecules. *Phys. Rev. Lett.*, 91(9), 2003. quant-ph/0301135.

[31] B. M. Terhal and D. P. DiVincenzo. Classical simulation of noninteracting-fermion quantum circuits. *Phys. Rev. A*, 65(032325), 2002. quant-ph/0108010.

[32] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

9 Appendix: Prediction Problems

In this appendix we prove a variant of Theorem 1.3 that is useful for prediction (as opposed to learning) problems—and that, as a bonus, is nearly tight. As usual, we first prove a general upper bound in terms of the fat-shattering dimension.

Theorem 9.1 *Let $\mathcal{S}$ be a sample space, let $\mathcal{C}$ be a p-concept class over $\mathcal{S}$, and let $\mathcal{D}$ be a probability measure over $\mathcal{S}$. Fix a p-concept $F \in \mathcal{C}$, as well an error parameter $\alpha > 0$. Suppose we are given m samples $X = (x_1, \ldots, x_m)$ drawn independently from $\mathcal{D}$, as well as bits $B = (b_1, \ldots, b_m)$ such that each b_i is 1 with independent probability $F(x_i)$. Suppose also that we choose a hypothesis $H \in \mathcal{C}$ to minimize $\sum_{i=1}^{m} |H(x_i) - b_i|$. Let*

$$\Delta_{H,F}(x) := H(x)(1 - F(x)) + (1 - H(x))F(x).$$

Then there exists a positive constant K such that

$$\mathop{\mathrm{EX}}_{x \in \mathcal{D}}[\Delta_{H,F}(x)] \leq \alpha + \inf_{C \in \mathcal{C}} \mathop{\mathrm{EX}}_{x \in \mathcal{D}}[\Delta_{C,F}(x)]$$

with probability at least $1 - \delta$ over X and B, provided that

$$m \geq \frac{K}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}}\left(\frac{\alpha}{10}\right)\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right).$$

Proof. As in the proof of Theorem 3.1, let $\mathcal{D}^*$ be the distribution over $(x, b) \in \mathcal{S} \times \{0, 1\}$ obtained by first drawing x from $\mathcal{D}$, then setting $b = 1$ with probability $F(x)$. Also, given a hypothesis $H \in \mathcal{C}$, let $H^*(x, 0) = H(x)$ and $H^*(x, 1) = 1 - H(x)$. Then let $\mathcal{C}^*$ be the p-concept class consisting of H^* for all $H \in \mathcal{C}$.

Call $H^* \in \mathcal{C}^*$ an α-good function if

$$\mathop{\mathrm{EX}}_{(x,b) \in \mathcal{D}^*}[H^*(x, b)] \leq \alpha + \inf_{C^* \in \mathcal{C}^*} \mathop{\mathrm{EX}}_{(x,b) \in \mathcal{D}^*}[C^*(x, b)].$$

Notice that if H^* is α-good, then

$$\mathop{\mathrm{EX}}_{x \in \mathcal{D}}[\Delta_{H,F}(x)] \leq \alpha + \inf_{C \in \mathcal{C}} \mathop{\mathrm{EX}}_{x \in \mathcal{D}}[\Delta_{C,F}(x)]$$

as desired.

Now, suppose we are given m samples $(x_1, b_1), \ldots, (x_m, b_m)$ drawn independently from $\mathcal{D}^*$. Then Theorem 2.3 implies that, if we choose $H^* \in \mathcal{C}^*$ to minimize

$$\sum_{i=1}^{m} H^*(x_i, b_i) = \sum_{i=1}^{m} |H(x_i) - b_i|,$$

then H^* will be α-good with probability at least $1 - \delta$, provided that

$$m = O\left(\frac{1}{\alpha^2}\left(\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\alpha}{5}\right)\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right).$$

Finally,

$$\mathrm{fat}_{\mathcal{C}^*}\left(\frac{\alpha}{5}\right) \leq \mathrm{fat}_{\mathcal{C}}\left(\frac{\alpha}{10}\right)$$

by the same argument as in Theorem 3.1. ■

Theorem 9.1 has the following immediate corollary.

Corollary 9.2 *Let ρ be an n-qubit state, let $\mathcal{D}$ be a distribution over two-outcome measurements, and let $\mathcal{E} = (E_1, \ldots, E_m)$ consist of m measurements drawn independently from $\mathcal{D}$. Suppose we are given bits $B = (b_1, \ldots, b_m)$, where each b_i is 1 with independent probability $\mathrm{Tr}(E_i\rho)$. Suppose also that we choose a hypothesis state σ to minimize $\sum_{i=1}^{m} |\mathrm{Tr}(E_i\sigma) - b_i|$. Let*

$$\Delta_{\sigma,\rho}(E) := \mathrm{Tr}(E\sigma)(1 - \mathrm{Tr}(E\rho)) + (1 - \mathrm{Tr}(E\sigma))\mathrm{Tr}(E\rho).$$

Then there exists a positive constant K such that

$$\underset{E\in\mathcal{D}}{\mathrm{EX}}\left[\Delta_{\sigma,\rho}\left(E\right)\right]\leq\alpha+\inf_{\varsigma}\underset{E\in\mathcal{D}}{\mathrm{EX}}\left[\Delta_{\varsigma,\rho}\left(E\right)\right]$$

with probability at least $1-\delta$ over $\mathcal{E}$ and B, provided that

$$m\geq\frac{K}{\alpha^{2}}\left(\frac{n}{\alpha^{2}}\log^{2}\frac{1}{\alpha}+\log\frac{1}{\delta}\right).$$

Let us describe a simple application of Corollary 9.2. Given a two-outcome measurement E and an n-qubit state ρ, let $E\left(\rho\right)\in\left\{0,1\right\}$ be the result of applying E to ρ—that is, $E\left(\rho\right)=1$ with probability $\mathrm{Tr}\left(E\rho\right)$ and $E\left(\rho\right)=0$ otherwise. Suppose our goal is to output a hypothesis state σ that maximizes $\Pr_{E\in\mathcal{D}}\left[E\left(\sigma\right)=E\left(\rho\right)\right]$, the "probability of agreement" between σ and ρ averaged over E. Corollary 9.2 shows that, by using $O\left(\frac{n}{\alpha^{4}}\log^{2}\frac{1}{\alpha}\right)$ measurements, we can get within an additive constant α of the maximum with high probability.

Similarly, suppose we are given a measurement E drawn from $\mathcal{D}$, and want to guess whether $E\left(\rho\right)$ will be 0 or 1. Here the maximum success probability is $\frac{1}{2}\underset{E\in\mathcal{D}}{\mathrm{EX}}\left[1+\left|2\,\mathrm{Tr}\left(E\rho\right)-1\right|\right]$, and is obtained by simply guessing 1 if $\mathrm{Tr}\left(E\rho\right)\geq\frac{1}{2}$, or 0 if $\mathrm{Tr}\left(E\rho\right)<\frac{1}{2}$. Again, it follows from Corollary 9.2 that by using $O\left(\frac{n}{\alpha^{4}}\log^{2}\frac{1}{\alpha}\right)$ measurements, we can get within an additive constant α of the maximum with high probability.

Using the same arguments as in Section 4, one can show that Corollary 9.2 is *tight* up to the $\log^{2}\frac{1}{\alpha}$ term—in particular, that

$$m=\Omega\left(\frac{1}{\alpha^{2}}\left(\frac{n}{\alpha^{2}}+\log\frac{1}{\delta}\right)\right)$$

measurements are needed. We omit the details.

What distinguishes this sort of prediction problem from the learning problems we have seen before is that, as the number of sample measurements m goes to infinity, we will *not* necessarily converge to the "true" state ρ. One way to see this is that, while ρ could be a mixed state, by convexity there is always a pure hypothesis state $\sigma=\left|\psi\right\rangle\left\langle\psi\right|$ that does as well at the prediction task as any other hypothesis. On the positive side, this means that to *find* such a hypothesis given the measurement results, it suffices to compute the principal eigenvector of a $2^{n}\times2^{n}$ matrix. Unlike for the learning problems, here there is no need for semidefinite or convex programming.

10 Appendix: More on YP

In Section 6.2, we defined the class YP, which consists of all problems solvable efficiently with the help of untrusted advice, and mentioned that this class might be of independent interest. Here we initiate the study of the structural complexity of YP, by proving four simple facts that relate YP to standard complexity classes.

Theorem 10.1

(i) $\mathsf{ZPP}\subseteq\mathsf{YP}$.

(ii) $\mathsf{YE}=\mathsf{NE}\cap\mathsf{coNE}$, *where* YE *is the exponential-time analogue of* YP *(i.e., both the advice size and the verifier's running time are* $2^{O(n)}$*).*

(iii) If $\mathsf{P}=\mathsf{YP}$ *then* $\mathsf{E}=\mathsf{NE}\cap\mathsf{coNE}$.

(iv) If $\mathsf{E}=\mathsf{NE}^{\mathsf{NP}^{\mathsf{NP}}}$ *then* $\mathsf{P}=\mathsf{YP}$.

Proof.

(i) Similar to the proof that $\mathsf{BPP}\subset\mathsf{P/poly}$. Given a ZPP machine M, first amplify M so that its failure probability on any input of length n is at most 2^{-2n}. Then by a counting argument, there exists a single random string r_{n} that causes M to succeed on all 2^{n} inputs simultaneously. Use that r_{n} as the YP machine's advice.

(ii) $\mathsf{YE} \subseteq \mathsf{NE} \cap \mathsf{coNE}$ is immediate. For $\mathsf{NE} \cap \mathsf{coNE} \subseteq \mathsf{YE}$, first concatenate the NE and coNE witnesses for all 2^n inputs of length n, then use the resulting string (of length $2^{O(n)}$) as the YE machine's advice.

(iii) If $\mathsf{P} = \mathsf{YP}$ then $\mathsf{E} = \mathsf{YE}$ by padding. Hence $\mathsf{E} = \mathsf{NE} \cap \mathsf{coNE}$ by part (ii).

(iv) Let M be a YP machine, and let y_n be the lexicographically first advice string that causes M to succeed on all 2^n inputs of length n. Consider the following computational problem: *given integers $\langle n, i \rangle$ encoded in binary, compute the i^{th} bit of y_n*. We claim that this problem is in $\mathsf{NE}^{\mathsf{NP}^{\mathsf{NP}}}$. For an $\mathsf{NE}^{\mathsf{NP}^{\mathsf{NP}}}$ machine can first guess y_n, then check that it works for all $x \in \{0,1\}^n$ using NP queries, then check that no lexicographically earlier string *also* works using $\mathsf{NP}^{\mathsf{NP}}$ queries, and finally return the i^{th} bit of y_n. So if $\mathsf{E} = \mathsf{NE}^{\mathsf{NP}^{\mathsf{NP}}}$, then the problem is in E, which means that an E machine can recover y_n itself by simply looping over all i. So if n and i take only logarithmically many bits to specify, then a P machine can recover y_n. Hence $\mathsf{P} = \mathsf{YP}$.

■

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/