The Learnability of Quantum States

Scott Aaronson*
MIT

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 one can do exponentially better in a statistical setting. In particular, to predict the outcomes of *most* measurements drawn from an arbitrary probability distribution, one only needs a number of sample measurements that grows linearly with n. This theorem has the conceptual implication that quantum states, despite being exponentially long vectors, are nevertheless "reasonable" in a learning theory sense. The theorem also has two applications to quantum computing: first, a new simulation of quantum one-way communication 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 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) [27]; to confirm the preparation of three-photon [26] and eight-ion [19] entangled states; to test controlled-NOT gates [24]; and to characterize optical devices [15].

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 in doing so. 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—Häffner et al. [19] report that, to reconstruct an entangled state of eight calcium ions, they needed to perform $656,100$ experiments! But to us it is a theoretical problem as well. For it suggests that learning an arbitrary state of (say) a thousand 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 such a state. 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 thereby lets us predict future observations!

Our purpose here is to propose a new resolution of this conundrum. We will show that, to predict the outcomes of "most" measurements on a quantum state, where "most" means with respect to any probability distribution of one's choice, it suffices to perform a number of sample measurements that grows only *linearly* with the number of qubits n. To be clear, this is not a replacement for standard quantum state tomography, since the hypothesis state that is output could be arbitrarily far from the true state in the usual trace distance metric. All we ask is that the hypothesis state is hard to distinguish from the true state with respect to a

*Email: aaronson@csail.mit.edu. This work was done while the author was a postdoc at the University of Waterloo, supported by CIAR through the Institute for Quantum Computing.

given distribution over measurements. This is a more modest goal—but even so, it might be surprising that changing the goal in this way gives an *exponential* improvement in the number of measurements required.

As a bonus, we will be able to use our learning theorem to prove two new results in quantum computing and information. The first result is a new relationship between randomized and quantum one-way communication complexities: namely that $\mathrm{R}^1(f) = O\left(M\,\mathrm{Q}^1(f)\right)$ for any partial or total Boolean function f, where $\mathrm{R}^1(f)$ is the randomized one-way communication complexity of f, $\mathrm{Q}^1(f)$ is the quantum one-way communication complexity, and M is the length of the recipient's input. The second result says that *trusted classical advice can be used to verify untrusted quantum advice on most inputs*—or in terms of complexity classes, that $\mathsf{HeurBQP/qpoly} \subseteq \mathsf{HeurQMA/poly}$. Both of these results follow from our learning theorem in intuitively-appealing ways; on the other hand, we would have no idea how to prove these results without the theorem.

We wish to stress that the main contribution of this paper is conceptual rather than technical. All of the 'heavy mathematical lifting' needed to prove the learning theorem has already been done: once one has the appropriate setup, the theorem follows readily by combining previous results due to Bartlett and Long [9] and Ambainis et al. [7]. Indeed, what is surprising to us is precisely that such a basic theorem was not discovered earlier.

The paper is organized as follows. We first give a formal statement of our learning theorem in Section 1.1, then answer objections to it in Section 1.2, situate it in the context of earlier work in Section 1.3, and discuss its implications in Section 1.4. In Section 2 we review some necessary results from computational learning theory and quantum information theory, and then prove our main theorem. Section 3 applies the learning theorem to communication complexity, while Section 4 applies it to quantum computational complexity and untrusted quantum advice. We conclude in Section 5 with some open problems.

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 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*

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

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

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

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

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

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

A proof will be given in Section 2.

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 a statistical one, 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 billions of years before we happen upon a measurement that distinguishes $|\psi\rangle$ from the maximally mixed state. Hence 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 needs the measurements to be drawn independently from some probability distribution $\mathcal{D}$. Is this not a strange assumption? Shouldn't one also allow adaptive measurements?*

Response. If all of our training data involved measurements in the $\{|0\rangle, |1\rangle\}$ basis, then regardless of how much data we had, clearly we couldn't hope to simulate a measurement in the $\{|+\rangle, |-\rangle\}$ basis. Therefore, as usual in learning theory, to get anywhere we need to make *some* assumption to the effect that the future will resemble the past. Such an assumption does not strike us as unreasonable in the context of quantum state estimation. For example, suppose that (as is often the case) the measurement process was itself stochastic, so that the experimenter did not know which observable was going to be measured until after it *was* measured. Or suppose the state was a "quantum program," which only had to succeed on typical inputs drawn from some probability distribution.[2]

However, with regard to the power of adaptive measurements, it is possible to ask somewhat more sophisticated questions. For example, suppose we perform a binary measurement E_1 (drawn from some distribution $\mathcal{D}$) on one copy of an n-qubit state ρ. Then, based on the outcome $z_1 \in \{0, 1\}$ of that measurement, suppose we perform another binary measurement E_2 (drawn from a new distribution $\mathcal{D}_{z_1}$) on a second copy of ρ; and so on for r copies of ρ. Finally, suppose we compute some Boolean function $f(z_1, \ldots, z_r)$ of the r measurement outcomes.

[2] At this point we should remind the reader that the distribution $\mathcal{D}$ over measurements only has to *exist*; it does not have to be *known*. All of our learning algorithms will be "distribution-free," in the sense that a single algorithm will work for any choice of $\mathcal{D}$.

Now, how many times will we need to repeat this adaptive procedure before, given $E_1, \ldots, E_r$ drawn as above, we can estimate (with high probability) the conditional probability that $f(z_1, \ldots, z_r) = 1$? If we simply apply Theorem 1.1 to the tensor product of all r registers, then it is easy to see that $O(nr)$ samples suffice. Furthermore, using ideas in the Electronic Supplementary Material, one can show that this is optimal: in other words, no improvement to (say) $O(n+r)$ samples is possible.

Indeed, even if we wanted to estimate the probabilities of all r of the measurement outcomes *simultaneously*, it follows from the union bound that we could do this with high probability, after a number of samples linear in n and polynomial in r.

We hope this illustrates how our learning theorem can be applied to more general settings than that for which it is explicitly stated. Naturally, there is a great deal of scope here for further research.

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 12 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 [18] that, if there exists a polynomial-time algorithm to find a Boolean circuit of size n consistent with observed data (whenever such a circuit exists), then there are no cryptographic one-way functions. Using the same techniques, one can show that, if there exists a polynomial-time quantum algorithm to prepare a state of n^k qubits consistent with observed data (whenever such a state exists), 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,[3] 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 quantum 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 the Electronic Supplementary Material, 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.*

[3] 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 less an objection to Theorem 1.1 than to quantum mechanics itself! With only one copy of ρ, the uncertainty principle immediately implies that not even statistical tomography is possible.

Objection 6 *One could never be certain that the condition of Theorem 1.1 was satisfied (in other words, that* $|\mathrm{Tr}(E_i\sigma) - \mathrm{Tr}(E_i\rho)| \leq \eta$ *for every* i*).*

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 apply each measurement only once, rather than multiple times? In that case, the above estimation strategy no longer works.*

Response. In the Electronic Supplementary Material, we prove a learning theorem that applies directly to this "measure-once" scenario. The disadvantage is that the upper bound on the number of measurements increases 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 $\sum_{i=1}^m (\mathrm{Tr}(E_i\sigma) - b_i)^2$. 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

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

Objection 8 *What if, instead of applying the "ideal" measurement E, the experimenter can only apply a noisy version E'?*

Response. If the noise that corrupts E to E' is governed by a known probability distribution such as a Gaussian, then E' is still just a POVM, so Theorem 1.1 applies directly. If the noise is adversarial, then we can also apply Theorem 1.1 directly, provided we have an upper bound on $|\mathrm{Tr}(E'\rho) - \mathrm{Tr}(E\rho)|$ (which simply gets absorbed into η).

Objection 9 *What if the measurements have $k > 2$ possible outcomes?*

Response. Here is a simple reduction to the two-outcome case. Before applying the k-outcome POVM $E = \left\{E^{(1)}, \ldots, E^{(k)}\right\}$, first choose an integer $j \in \{1, \ldots, k\}$ uniformly at random, and then pretend that the POVM being applied is $\left\{E^{(j)}, I - E^{(j)}\right\}$ (i.e., ignore the other $k-1$ outcomes). By the union bound, if our goal is to ensure that

$$\Pr_{E\in\mathcal{D}}\left[\sum_{j=1}^k \left|\mathrm{Tr}\left(E^{(j)}\sigma\right) - \mathrm{Tr}\left(E^{(j)}\rho\right)\right| > \gamma\right] \leq \varepsilon$$

with probability at least $1-\delta$, then in our upper bounds it suffices to replace every occurrence of γ by γ/k, and every occurrence of ε by ε/k. We believe that one could do better than this by analyzing the k-outcome case directly; we leave this as an open problem.[4]

[4]Notice that any sample complexity bound must have at least a linear dependence on k. Here is a proof sketch: given a subset $S \subseteq \{1, \ldots, k\}$ with $|S| = k/2$, let $|S\rangle$ be a uniform superposition over the elements of S. Now consider simulating a

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

Computational learning theory can be understood as a modern response to David Hume's Problem of Induction: "if an ornithologist sees 500 ravens and all of them are black, why does that provide any grounds at all for expecting the 501st raven to be black? After all, the hypothesis that the 501st raven will be white seems 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 1989 paper, Blumer et al. [11] 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. Indeed, 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, these results apply only to Boolean hypothesis classes. So to prove our learning theorem, we will need a more powerful result due to Bartlett and Long [9], which upper-bounds the number of data points needed to learn *real*-valued hypothesis classes.

Quantum Information Theory

Besides results from classical learning theory, we will also need a result of Ambainis et al. [7] in quantum information theory. Ambainis et al. showed 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 need $n \geq (1 - H(p))\,k$, where H is the binary entropy function.

Perhaps the central idea of this paper is to turn Ambainis et al.'s result on its head, and see it not as lower-bounding the number of qubits needed for coding and communication tasks, but instead as *upper*-bounding the "effective dimension" of a quantum state to be learned. (In theoretical computer science, this is hardly the first time that a negative result has been turned into a positive one. A similar "lemons-into-lemonade" conceptual shift was made by Linial, Mansour, and Nisan [23], when they used a limitation of constant-depth circuits to give an efficient algorithm for learning those circuits.)

Quantum State Estimation

Physicists have been interested in quantum state estimation since at least the 1950's (see [25] for a good overview). For practical reasons, they have been particularly concerned with minimizing the number of measurements. However, most 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 body of work on how to estimate a quantum state given incomplete measurement results—see Bužek et al. [14] for a good introduction to the subject, or Bužek [13] for estimation algorithms that are similar in spirit to ours. But there are at least two differences between the previous work and ours. First, while some of the previous work offers numerical evidence that few measurements seem to suffice in practice, so far as we know none of it considers asymptotic complexity. Second, the previous work almost always assumes that an experimenter starts with a prior probability distribution over quantum states (often the uniform distribution), and then either updates the distribution using Bayes' rule, or else applies a Maximum-Likelihood principle. By contrast, our learning approach requires no assumptions about a distribution over states; it instead requires only a (possibly-unknown) distribution over *measurements*. The advantage of the latter approach, in our view, is that an experimenter has much more control over which measurements to apply than over the nature of the state to be learned.

measurement of $|S\rangle$ in the computational basis, $\{|1\rangle, \ldots, |k\rangle\}$. It is clear that $\Omega(k)$ sample measurements are needed to do this even approximately.

1.4 Implications

The main implication of our learning theorem is conceptual: it shows that quantum states, considered as a hypothesis class, are "reasonable" in the sense of computational learning theory. Were this *not* the case, it would presumably strengthen the view of quantum computing skeptics [17, 22] that quantum states are "inherently extravagant" objects, which will need to be discarded as our knowledge of physics expands. (Or at least, it would suggest that the "operationally meaningful" quantum states comprise only a tiny portion of Hilbert space.) 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 could be of practical use in quantum state estimation, since it provides 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 result is not directly applicable, we hope the mere *fact* that this sort of learning is possible will serve as a spur to further research. As an analogy, classical computational learning theory has had a large influence on neural networks, computer vision, and other fields,[5] but this influence might have had less to do with the results themselves than with their philosophical moral.

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 $f(x, y)$ with high probability? Note that there is no back-communication from Bob to Alice.

Let $\mathrm{R}^1(f)$, and $\mathrm{Q}^1(f)$ be the number of bits that Alice needs to send, if her message to Bob is randomized or quantum respectively.[6] Then improving an earlier result of Aaronson [1], in Section 3 we are able to show the following:

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.*

Intuitively, this means that if Bob's input is small, then quantum communication provides at most a small advantage over classical communication.

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 *approximate verification of quantum software.* Suppose you want to evaluate some Boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$, on typical inputs x drawn from a probability distribution $\mathcal{D}$. So you go to the quantum software store and purchase $|\psi_f\rangle$, a q-qubit piece of quantum software. The software vendor tells you that, to evaluate $f(x)$ on any given input $x \in \{0,1\}^n$, you simply need to apply a fixed measurement E to the state $|\psi_f\rangle|x\rangle$. However, you do not trust $|\psi_f\rangle$ to work as expected. Thus, the following question arises: is there a fixed, polynomial-size set of "benchmark inputs" $x_1, \ldots, x_T$, such that for *any* quantum program $|\psi_f\rangle$, if $|\psi_f\rangle$ works on the benchmark inputs then it will also work on most inputs drawn from $\mathcal{D}$?

Using our learning theorem, we will show in Section 4 that the answer is yes. Indeed, we will actually go further than that, and give an *efficient procedure* to test $|\psi_f\rangle$ against the benchmark inputs. The central difficulty here is that the measurements intended to test $|\psi_f\rangle$ might also destroy it. We will resolve this difficulty by means of a "Witness Protection Lemma," which might have applications elsewhere.

In terms of complexity classes, we can state our verification theorem as follows:

[5] According to Google Scholar, Valiant's original paper on the subject [29] has been cited 1918 times as of this writing, with a large fraction of the citations coming from fields other than theoretical computer science.

[6] Here the superscript '1' denotes one-way communication.

Theorem 1.5 $\mathsf{HeurBQP/qpoly} \subseteq \mathsf{HeurQMA/poly}$.

Here $\mathsf{BQP/qpoly}$ is the class of problems solvable in quantum polynomial time, with help from a polynomial-size "quantum advice state" $|\psi_n\rangle$ that depends only on the input length n; while QMA (Quantum Merlin-Arthur) is the class of problems for which a 'yes' answer admits a polynomial-size quantum proof. Then $\mathsf{HeurBQP/qpoly}$ and $\mathsf{HeurQMA/poly}$ are the *heuristic* versions of $\mathsf{BQP/qpoly}$ and $\mathsf{QMA/poly}$ respectively—that is, the versions where we only want to succeed on most inputs rather than all of them.

2 The Measurement Complexity of Quantum Learning

We now 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 Ambainis et al. [7] 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 "Occam's Razor Theorem" of Blumer et al. [11], which is stated in terms of a parameter called VC dimension. However, Blumer et 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 [20]—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 as straightforward as one might hope. Fortunately, various authors [6, 8, 9, 10, 20] have already done most of the work for us, with results due to Anthony and Bartlett [8] and to Bartlett and Long [9] being particularly relevant. To state their results, we 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 $F : \mathcal{S} \to [0,1]$, and a *p-concept class over* $\mathcal{S}$ is a set of p-concepts over $\mathcal{S}$. Kearns and Schapire [20] 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 there is no finite such maximum, then $\mathrm{fat}_{\mathcal{C}}(\gamma) = \infty$.)

We can now state the result of Anthony and Bartlett.

Theorem 2.2 (Anthony and Bartlett [8]) *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.[7]

Theorem 2.3 (Bartlett and Long [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 a p-concept $F : \mathcal{S} \to [0,1]$ (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*

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

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.

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

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

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

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

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

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

[7] 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 $\mathcal{C}$.

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

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

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 $F_\rho : \mathcal{S} \to [0,1]$ 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 Ambainis et al. [7], which upper-bounds the number of classical bits that can be "encoded" into n qubits.

Theorem 2.5 (Ambainis et al. [7]) *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 $n/\gamma^2 = \Omega(k)$.[8]

■

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

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

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

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

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

3 Application to Quantum Communication

In this section we use our quantum learning theorem to prove a new result about *one-way communication complexity*. Here 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 $a \in \mathcal{D}_x$.[9] 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 $\mathrm{R}^1(f) = \mathrm{Q}^1(f) = O(\log N)$.[10] Furthermore, Gavinsky et al. [16] have recently shown that there exists a partial function f for which $\mathrm{R}^1(f) = \Omega\left(\sqrt{N}\right)$ but $\mathrm{Q}^1(f) = O(\log N)$.

On the other hand, it follows from a result of Klauck [21] 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 [1] 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 an open problem for several years.

Using our quantum learning theorem, we are able to resolve this problem, 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.

[8] 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 Ambainis et al. [7]. Doing so, we obtain $n/\gamma^2 \geq 2k/\ln 2$.

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

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

Proof of Theorem 1.4. Let $f : \mathcal{Z} \to \{0,1\}$ 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 $\mathcal{D}$.[11]

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 $|\mathrm{Tr}(E_{y_i}\sigma) - f(x, y_i)| \leq \eta$ 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 $|\mathrm{Tr}(E_{y_i}\sigma) - \mathrm{Tr}(E_{y_i}\rho_x)| \leq \eta$ 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,

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

with probability at least $1 - \delta$ 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

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

with probability at least $1 - \delta$. 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[12] yields a total f for which $\mathrm{R}^1(f)$ is exponentially larger than M, while the recent results of Gavinsky et al. [16] 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

(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 $\ell \in \{1, \ldots, \alpha N\}$.

[11] 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.

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

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 $\alpha > 0$.[13] 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* $\alpha \ll 1/\sqrt{\log N}$.[14] 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 $\mathrm{D}^1(f) = O\left(M\,\mathrm{Q}^1(f)\log \mathrm{Q}^1(f)\right)$ [1] is *tight* up to a polylogarithmic term. This answers one of the open questions in [1].

4 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 $L : \{0,1\}^* \to \{0,1\}$ 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.

4.1 Quantum Advice and Proofs

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 4.1 *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 [1] 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,[15] 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).

[13] 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.

[14] R. de Wolf, personal communication.

[15] See [2] for a detailed definition, as well as a proof that PostBQP coincides with the classical complexity class PP.

Definition 4.2 *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 $|\varphi\rangle \in \mathcal{H}_2^{\otimes \operatorname{poly}(n)}$ 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.

4.2 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:[16]

Definition 4.3 *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 $\mathsf{P/poly}$ and in $\mathsf{NP} \cap \mathsf{coNP}$. Indeed, while we are at it, let us initiate the study of YP, by mentioning four simple facts that relate YP to standard complexity classes.

Theorem 4.4

(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).

[16] 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'.

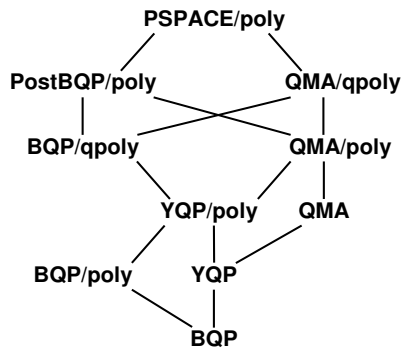

Figure 1: Some quantum advice and proof classes. The containment $\mathsf{BQP/qpoly} \subseteq \mathsf{PostBQP/poly}$ was shown in [1], while $\mathsf{QMA/qpoly} \subseteq \mathsf{PSPACE/poly}$ was shown in [3].

(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 ith 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 ith 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 $\mathsf{P} = \mathsf{YP}$.

■

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 4.5 *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 $|\varphi_n\rangle \in \mathcal{H}_2^{\otimes \operatorname{poly}(n)}$ 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 $\mathsf{BQP/qpoly}$ and in $\mathsf{QMA} \cap \mathsf{coQMA}$. We also have $\mathsf{YQP/qpoly} = \mathsf{BQP/qpoly}$, since the untrusted YQP advice can be tacked onto the trusted $/\mathsf{qpoly}$ advice. Figure 1 shows the known containments among various classes involving quantum advice and proofs.

4.3 Heuristic Complexity

Ideally, we would like to show that $\mathsf{BQP/qpoly} = \mathsf{YQP/poly}$—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.[17] We now explain what this means (for details, see the excellent survey by Bogdanov and Trevisan [12]).

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

[17] 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.

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 $1-\frac{1}{\mathrm{poly}(n)}$ fraction of inputs.

Definition 4.6 *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 4.7 *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 4.8 *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 $|\varphi_{n,\varepsilon}\rangle \in \mathcal{H}_2^{\otimes \mathrm{poly}(n,1/\varepsilon)}$ 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 $P_A^{1-L(x)}\left(|x\rangle |\varphi\rangle\right) \geq \frac{1}{3}$ is at most ε.

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

The following table summarizes the most important complexity classes, prefixes, and suffixes defined in the sections above.

BQP	Bounded-Error Quantum Polynomial-Time
QMA	Quantum Merlin-Arthur (BQP with a quantum witness depending on input x)
YQP	Yoda Quantum Polynomial-Time (BQP with a quantum witness independent of x)
/poly	With polynomial-size classical advice
/qpoly	With polynomial-size quantum advice
Post	With postselection
Heur	Heuristic (only needs to work for most inputs)

4.4 Proof

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 intuition, we need the following "quantum union bound," which was proved by Aaronson [3] based on a result of Ambainis et al. [7].

Proposition 4.9 (Aaronson [3]) *Let $E_1, \ldots, E_m$ be two-outcome measurements, and suppose $\mathrm{Tr}(E_i\rho) \geq 1-\epsilon$ for all $i \in \{1, \ldots, m\}$. Then if we apply $E_1, \ldots, E_m$ 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 4.9, we can prove the following "Witness Protection Lemma."

Lemma 4.10 (Witness Protection Lemma) *Let $\mathcal{E} = \{E_1, \ldots, E_m\}$ 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 4.9. For property (iii), let ρ_u be the state of ρ immediately after the uth iteration, conditioned on iterations $1, \ldots, u$ 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 $1, \ldots, u$. 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

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

■

Finally, by using Lemma 4.10, we can prove Theorem 1.5: that $\mathsf{HeurBQP/qpoly} = \mathsf{HeurYQP/poly} \subseteq \mathsf{HeurQMA/poly}$.

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

(c) If $P_{A^*}^{L(x_i)}(|x_i\rangle |\varphi\rangle) \geq \frac{17}{18}$ 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 4.10. Setting $\epsilon := e^{-\ell/18}$, we have $P_{A^*}^{L(x_i)}\left(|x_i\rangle |\psi_{n,\varepsilon}\rangle^{\otimes \ell}\right) \geq 1-\epsilon$ 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 4.10. Set $\lambda := \frac{1}{3}$. Then if Phase 1 succeeds with probability at least λ, for all i we have

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

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

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

and also

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

Hence

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

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

5 Summary and 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 [4] or noninteracting-fermion states [28]?[18]

[18] 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 our statistical quantum state learning approach in photonics, ion traps, NMR, or any other technology that allows the preparation and measurement of multi-qubit entangled states. Already for three or four 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. How much of an improvement partly depends on how far our learning results can be improved, as well as on what the constant factors are. A related issue is that, while one can always reduce noisy, k-outcome measurements to the noiseless, two-outcome measurements that we consider, one could almost certainly prove better upper bounds by analyzing realistic measurements more directly.

One might hope for a far-reaching generalization of our learning theorem, 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.[19] Still, it would be interesting to know if there is anything to say about statistical 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 [5]) 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.

6 Acknowledgments

I thank Noga Alon, Dave Bacon, Peter Bartlett, Robin Blume-Kohout, Andy Drucker, Aram Harrow, Tony Leggett, Peter Shor, Luca Trevisan, Umesh Vazirani, Ronald de Wolf, and the anonymous reviewers for helpful discussions and correspondence.

References

[1] S. Aaronson. Limitations of quantum advice and one-way communication. *Theory of Computing*, 1:1–28, 2005. quant-ph/0402095. Conference version in Proceedings of CCC'2004.

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

[3] 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.

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

[5] S. Aaronson and G. Kuperberg. Quantum versus classical proofs and advice. To appear in Theory of Computing. quant-ph/0604056, 2006.

[6] 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.

[7] 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.

[19] 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 $f : \{0,1\}^{n-1} \to \{0,1\}$. 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.

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

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

[10] 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.

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

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

[13] V. Bužek. Quantum tomography from incomplete data via MaxEnt principle. In [25]. Pages 189-234.

[14] 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.

[15] 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.

[16] D. Gavinsky, J. Kempe, I. Kerenidis, R. Raz, and R. de Wolf. Exponential separations for one-way quantum communication complexity, with applications to cryptography. In *Proc. ACM STOC*, 2007. To appear. quant-ph/0611209.

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

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

[19] H. Häffner, W. Hänsel, C. F. Roos, J. Benhelm, D. Chek-al-kar, M. Chwalla, T. Körber, U. D. Rapol, M. Riebe, P. O. Schmidt, C. Becher, O. Gühne, W. Dür, and R. Blatt. Scalable multiparticle entanglement of trapped ions. *Nature*, 438:643–646, 2005. quant-ph/0603217.

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

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

[22] 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.

[23] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

[24] 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.

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

[26] 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.

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

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

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