Shadow Tomography of Quantum States*

Scott Aaronson†

Abstract

We introduce the problem of *shadow tomography*: given an unknown D-dimensional quantum mixed state ρ, as well as known two-outcome measurements $E_1, \ldots, E_M$, estimate the probability that E_i accepts ρ, to within additive error ε, for each of the M measurements. How many copies of ρ are needed to achieve this, with high probability? Surprisingly, we give a procedure that solves the problem by measuring only $\widetilde{O}\left(\varepsilon^{-4} \cdot \log^4 M \cdot \log D\right)$ copies. This means, for example, that we can learn the behavior of an arbitrary n-qubit state, on *all* accepting/rejecting circuits of some fixed polynomial size, by measuring only $n^{O(1)}$ copies of the state. This resolves an open problem of the author, which arose from his work on private-key quantum money schemes, but which also has applications to quantum copy-protected software, quantum advice, and quantum one-way communication. Recently, building on this work, Brandão et al. have given a different approach to shadow tomography using semidefinite programming, which achieves a savings in computation time.

1 Introduction

One of the most striking features of quantum mechanics is the *destructive nature of measurement.* Given a single copy of a quantum state ρ, which is otherwise unknown to us, no amount of cleverness will ever let us recover a classical description of ρ, even approximately, by measuring ρ. Of course, the destructive nature of measurement is what opens up many of the cryptographic possibilities of quantum information, including quantum key distribution and quantum money.

In general, the task of recovering a description of a D-dimensional quantum mixed state ρ, given many copies of ρ, is called *quantum state tomography.* This task can be shown for information-theoretic reasons to require $\Omega\left(D^2\right)$ copies of ρ, while a recent breakthrough of O'Donnell and Wright [27] and Haah et al. [18] showed that $O\left(D^2\right)$ copies also suffice.[1] Unfortunately, this number can be astronomically infeasible: recall that, if ρ is a state of n entangled qubits, then $D = 2^n$. No wonder that the world record, for full[2] quantum state tomography, is 10-qubit states, for which millions of measurements were needed [28].

Besides the practical issue, this state of affairs could be viewed as an epistemic problem for quantum mechanics itself. If learning a full description of an n-qubit state ρ requires measuring

*An earlier version of this paper appeared in the Proceedings of the 2018 ACM Symposium on Theory of Computing (STOC), pages 325-338

†University of Texas at Austin. Email: aaronson@cs.utexas.edu. Supported by a Vannevar Bush Fellowship from the US Department of Defense, a Simons Investigator Award, and the Simons "It from Qubit" collaboration.

[1] Here and throughout this paper, the notations $\widetilde{\Omega}$ and $\widetilde{O}$ mean we suppress polylogarithmic factors.

[2] By "full," we mean that the procedure could have recovered the state ρ regardless of what it was, rather than requiring an assumption like, e.g., that ρ has a small matrix product state description.

$\exp(n)$ copies of ρ, then should we even say that the full description is "there" at all, in a single copy of ρ?

Naturally, we could ask the same question about a classical probability distribution $\mathcal{D}$ over n-bit strings. In that case, the exponentiality seems to militate toward the view that no, the vector of 2^n probabilities is *not* "out there" in the world, but is only "in our heads," while what's "out there" are just the actual n-bit samples from $\mathcal{D}$, along with whatever physical process generated the samples. Quantum mechanics is different, though, because 2^n amplitudes can interfere with each other: an observable effect that seems manifestly *not* just in our heads! Interference forces us to ask the question anew.

Partly inspired by these thoughts, a long line of research has sought to show that, once we impose some reasonable operational restrictions on what a quantum state will be used for, an n-qubit state ρ actually contains "much less information than meets the eye": more like n or $n^{O(1)}$ classical bits than like 2^n bits. Perhaps the "original" result along these lines was Holevo's Theorem [21], which says that by sending an n-qubit state, Alice can communicate at most n classical bits to Bob (or $2n$, if Alice and Bob have pre-shared entanglement). Subsequently, the random access code lower bound of Ambainis, Nayak, Ta-Shma, and Vazirani [12] showed that this is still true, even if Bob wants to learn just a single one of Alice's bits.

Since 2004, a series of results by the author and others has carried the basic conclusion further. Very briefly, these results have included the postselected learning theorem [1]; the Quantum Occam's Razor Theorem [4]; the "de-Merlinization" of quantum protocols [3]; a full characterization of quantum advice [10]; and a recent online learning theorem for quantum states [8]. We'll apply tools from several of those results in this paper, and will discuss the results later in the introduction where it's relevant. In any case, though, none of the previous results directly addressed the question: *information-theoretically, how much can be learned about an n-qubit state ρ by measuring only $n^{O(1)}$ copies of ρ?*

1.1 Our Result

Motivated by the above question, this paper studies a basic new task that we call *shadow tomography*, and define as follows.

Problem 1 (Shadow Tomography) *Given an unknown D-dimensional quantum mixed state ρ, as well as known 2-outcome measurements $E_1, \ldots, E_M$, each of which accepts ρ with probability $\operatorname{Tr}(E_i \rho)$ and rejects ρ with probability $1 - \operatorname{Tr}(E_i \rho)$, output numbers $b_1, \ldots, b_M \in [0,1]$ such that $|b_i - \operatorname{Tr}(E_i \rho)| \leq \varepsilon$ for all i, with success probability at least $1 - \delta$. Do this via a measurement of $\rho^{\otimes k}$, where $k = k(D, M, \varepsilon, \delta)$ is as small as possible.*

The name "shadow tomography" was suggested to us by Steve Flammia, and refers to the fact that we aim to recover, not the full density matrix of ρ, but only the "shadow" that ρ casts on the measurements $E_1, \ldots, E_M$.

Observe, for a start, that shadow tomography is easy to achieve using $k = O\left(D^2/\varepsilon^2\right)$ copies of the state ρ, by just ignoring the E_i's and doing full tomography on ρ, using the recent protocols of O'Donnell and Wright [27] or Haah et al. [18]. At a different extreme of parameters, shadow tomography is *also* easy using $k = \widetilde{O}\left(M/\varepsilon^2\right)$ copies of ρ, by just applying each measurement E_i to separate copies of ρ.

At a mini-course taught in February 2016 (see [6, Section 8.3.1]), the author discussed shadow tomography—though without calling it that—and posed the question, *what happens if D and*

M are both exponentially large? Is it conceivable that, even then, the M expectation values $\operatorname{Tr}(E_1\rho), \ldots, \operatorname{Tr}(E_M\rho)$ could all be approximated using only, say, $\operatorname{poly}(\log D, \log M)$ copies of the state ρ? The author didn't venture to guess an answer; other researchers' opinions were also divided.

The main result of this paper is to settle the question affirmatively.

Theorem 2 (Shadow Tomography Theorem) *Problem 1 (Shadow Tomography) is solvable using only*

$$k = \widetilde{O}\left(\frac{\log 1/\delta}{\varepsilon^4} \cdot \log^4 M \cdot \log D\right)$$

copies of the state ρ, where the $\widetilde{O}$ hides a $\operatorname{poly}\left(\log\log M, \log\log D, \log\frac{1}{\varepsilon}\right)$ *factor.*[3] *The procedure is fully explicit.*

In Section 1.2, we'll give an overview of the proof of this theorem. In Section 2, we'll discuss the motivation, and give applications to quantum money, quantum copy-protected software, quantum advice, and quantum one-way communication. For now, let's make some initial comments about the theorem itself: why it's nontrivial, why it's consistent with other results, etc.

The key point is that Theorem 2 lets us learn the behavior of a state of exponential dimension, with respect to exponentially many different observables, using only polynomially many copies of the state. To achieve this requires measuring the copies in an extremely careful way, to avoid destroying them as we proceed.

Naturally, to implement the required measurement on $\rho^{\otimes k}$ could, in the worst case, require a quantum circuit of size polynomial in both M and D. (Note that the input—i.e., the list of measurement operators E_i—already involves $\Theta\left(MD^2\right)$ complex numbers.) It's interesting to study how much we can improve the *computational* complexity of shadow tomography, with or without additional assumptions on the state ρ and measurements E_i. In Sections 1.3 and 7, we'll say more about this question, and about recent work by Brandão et al. [14], which builds on our work to address it. In this paper, though, our main focus is on the information-theoretic aspect, of how many copies of ρ are needed.

The best *lower* bound that we know on the number of copies is $\Omega\left(\frac{\min\left\{D^2, \log M\right\}}{\varepsilon^2}\right)$.[4] We'll prove this lower bound in Section 6, using an information theory argument. We'll also observe that a lower bound of $\Omega\left(\frac{\min\{D, \log M\}}{\varepsilon^2}\right)$ holds even in the special case where the state and measurements are entirely classical—in which case the lower bound is actually *tight*. In the general (quantum) case, we don't know whether shadow tomography might be possible using $(\log M)^{O(1)}$ copies, independent of the Hilbert space dimension D.

But stepping back, why isn't even Theorem 2 immediately ruled out by, for example, Holevo's Theorem [21]—which says (roughly) that by measuring a D-dimensional state, we can learn at most $O(\log D)$ independent classical bits? One way to answer this question is to observe that there's

[3] In an earlier version of this paper, the dependence on ε was $1/\varepsilon^5$. The improvement to $1/\varepsilon^4$ comes from using the recent online learning algorithm of Aaronson et al. [8].

[4] In an earlier version of this paper, we proved only a weaker lower bound, namely $\Omega\left(\frac{\log M}{\varepsilon^2}\right)$ assuming D can be arbitrarily large. In this version, we've reworked the lower bound to incorporate the dependence on D explicitly. One side effect is that, if we let M be arbitrarily large, then our lower bound now subsumes the $\Omega\left(D^2\right)$ lower bound on the sample complexity of *ordinary* quantum state tomography, originally proved by O'Donnell and Wright [27] and Haah et al. [18].

no claim that the M numbers $\operatorname{Tr}(E_i\rho)$ can all be varied independently of each other by varying ρ: indeed, it follows from known results [12, 4] that they can't be, unless $D = \exp(\Omega(M))$.

Another answer is as follows. It's true that there exist so-called *tomographically complete* sets of two-outcome measurements, of size $M = O(D^2)$. These are sets $E_1, \ldots, E_M$ such that knowing $\operatorname{Tr}(E_i\rho)$ exactly, for every $i \in [M]$, suffices to determine ρ itself. So if we ran our shadow tomography procedure on a tomographically complete set, with small enough ε, then we could reconstruct ρ, something that we know requires $k = \Omega(D^2)$ copies of ρ. However, this would require knowing the $\operatorname{Tr}(E_i\rho)$'s to within additive error $\varepsilon \ll 1/D$, which remains perfectly compatible with a shadow tomography procedure that uses $\operatorname{poly}(\log M, \log D, \varepsilon^{-1})$ copies.

One last clarifying remark is in order. After satisfying themselves that it's not impossible, some readers might wonder whether Theorem 2 follows trivially from the so-called "Gentle Measurement Lemma" [32, 1], which is closely related to the concept of *weak measurement* in physics. We'll explain gentle measurement in more detail in Section 3, but loosely speaking, the idea is that *if* the outcome of a measurement E on a state ρ could be predicted almost with certainty, given knowledge of ρ, *then* E can be implemented in a way that damages ρ very little, leaving the state available for future measurements. Gentle measurement will play an important role in the proof of Theorem 2, as it does in many quantum information results.

However, all that we can *easily* deduce from gentle measurement is a "promise-gap" version of Theorem 2. In particular: suppose we're given real numbers $c_1, \ldots, c_M \in [0,1]$, and are promised that for each $i \in [M]$, either $\operatorname{Tr}(E_i\rho) \geq c_i$ or $\operatorname{Tr}(E_i\rho) \leq c_i - \varepsilon$. In that case, we'll state and prove, as Proposition 20, that it's possible to decide which of these holds, for every $i \in [M]$, with high probability using only $k = O\left(\frac{\log M}{\varepsilon^2}\right)$ copies of ρ. This is because, *given the promise gap*, we can design an "amplified" version of E_i that decides which side of the gap we're on while damaging $\rho^{\otimes k}$ only very little.

But what if there's no promise, as there typically isn't in real-world tomography problems? In that case, the above approach fails utterly: indeed, every two-outcome measurement E that we could possibly apply seems dangerous, because if ρ happens to be "just on the knife-edge" between acceptance and rejection—a possibility that we can never rule out—then applying E to copies of ρ will severely damage those copies. And while we can afford to lose a *few* copies of ρ, we have only $\operatorname{poly}(\log M, \log D)$ copies in total, which is typically far fewer than the M measurement outcomes that we need to learn.[5] This is the central problem that we solve.

1.2 Techniques

At a high level, our shadow tomography procedure involves combining two ideas.

The first idea is *postselected learning of quantum states*. This tool was introduced by Aaronson [1] in 2004 to prove the complexity class containment $\mathsf{BQP/qpoly} \subseteq \mathsf{PostBQP/poly}$, where $\mathsf{BQP/qpoly}$ means BQP augmented with polynomial-size quantum advice, and $\mathsf{PostBQP}$ means BQP augmented with postselected measurements, a class that equals PP by another result of Aaronson

[5] As an alternative, one might hope to prove Theorem 2 by simply performing a series of "weak measurements" on the state $\rho^{\otimes k}$, which would estimate the real-valued observables $\operatorname{Tr}(E_i\rho)$, but with Gaussian noise of variance $\gg 1/k$ deliberately added to the measurement outcomes, in order to prevent $\rho^{\otimes k}$ from being damaged too much by the measurements. However, a calculation reveals that every such measurement could damage the state by $1/k^{O(1)}$ in variation distance. Thus, while this strategy would let us safely estimate $\operatorname{poly}(\log M, \log D)$ observables $\operatorname{Tr}(E_i\rho)$ in succession, it doesn't appear to let us estimate all M of them.

[2]. Postselected learning is related to *boosting* in computational learning theory, as well as to the multiplicative weights update method.

Restated in the language of this paper, the canonical example of postselected learning is as follows. Suppose Alice knows the complete classical description of a D-dimensional quantum mixed state ρ, and suppose she wants to describe ρ to Bob over a classical channel—well enough that Bob can approximate the value of $\operatorname{Tr}(E_i\rho)$, for each of M two-outcome measurements $E_1,\ldots,E_M$ known to both players. To do this, Alice could always send over the full classical description of ρ, requiring $\Theta\left(D^2\right)$ bits. Or she could send the values of the $\operatorname{Tr}(E_i\rho)$'s, requiring $\Theta(M)$ bits.

But there's also something much more efficient that Alice can do, requiring only $\Theta(\log D\cdot\log M)$ bits. Namely, she can assume that, being totally ignorant at first, Bob's "initial guess" about ρ is simply that it's the maximally mixed state, $\rho_0:=\frac{I}{D}$. She can then repeatedly help Bob to refine his current guess ρ_t to a better guess ρ_{t+1}, by telling Bob the index i of a measurement on which his current guess badly fails—that is, on which $|\operatorname{Tr}(E_i\rho_t)-\operatorname{Tr}(E_i\rho)|$ is large—as well as the approximate value of $\operatorname{Tr}(E_i\rho)$. To use this information, Bob can let ρ_{t+1} be the state obtained by starting from ρ_t (or technically, an amplified version of ρ_t), measuring the observable E_i, and then *postselecting* (that is, conditioning) on getting measurement outcomes that are consistent with ρ. Of course this postselection might have only a small chance of success, were Bob doing it with the actual state ρ_t, but he can instead *simulate* postselection using a classical description of ρ_t.

The key question, with this approach, is how many iterations T are needed until Bob converges to a hypothesis state ρ_T such that $\operatorname{Tr}(E_i\rho_T)\approx\operatorname{Tr}(E_i\rho)$ for every i. And the key result is that only $\Theta(\log D)$ iterations are needed. Intuitively, this is because the ground truth, ρ, has "weight" at least $\frac{1}{D}$ within the maximally mixed state $\frac{I}{D}$. Repeatedly choosing measurements where the current hypothesis still does poorly, and then postselecting on doing well on those measurements, causes all the components of $\frac{I}{D}$ *other than* ρ to decay at an exponential rate, until a measurement can no longer be found where the current hypothesis does poorly. That might happen well before we reach ρ itself, but if not, then ρ itself will be reached after $\Theta(\log D)$ iterations.

Postselected learning has since found further uses in quantum computing theory [3, 8]. But there seems to be a fundamental difficulty in applying it to shadow tomography. Namely, in shadow tomography *there's no "Alice"*: that is, no agent who knows a classical description of the state ρ, and who can thus helpfully point to measurements E_i that are useful for learning ρ's behavior. So any shadow tomography procedure will need to find informative measurements by itself, and do so using only polylogarithmically many copies of ρ.

The second idea, the *gentle search procedure*, does exactly that. In 2006, as a central ingredient in the proof of the complexity class containment $\mathsf{QMA/qpoly}\subseteq\mathsf{PSPACE/poly}$, Aaronson [3] claimed a result that he called "Quantum OR Bound." This result can be stated as follows: given an unknown state ρ and known two-outcome measurements $E_1,\ldots,E_M$, there is a procedure, using $k=O\left(\frac{\log M}{\varepsilon^2}\right)$ copies of ρ, to decide whether

(i) some E_i accepts ρ with probability at least c or

(ii) no E_i accepts ρ with probability greater than $c-\varepsilon$,

with high probability and assuming one of the cases holds. Note that the number of copies is not only logarithmic in M, but independent of the dimension of ρ.

Aaronson's proof of the Quantum OR Bound was based on simply applying amplified versions of the E_i's to $\rho^{\otimes k}$ in a random order, and checking whether any of the measurements accepted.

Unfortunately, Aaronson's proof had an error, which was discovered in 2016 by Harrow, Lin, and Montanaro [19]. Happily, Harrow et al. also fixed the error, thereby recovering all the consequences that Aaronson had claimed, as well as new consequences. To do so, Harrow et al. designed two new measurement procedures, both of which solve the problem: one based on the "in-place amplification" of Marriott and Watrous [23], and another that applies amplified E_i's conditional on a control qubit being $|1\rangle$, and that checks not only whether any of the measurements accept but also whether the control qubit has decohered. It remains open whether Aaronson's original procedure is also sound.

For shadow tomography, however, there's a further problem. Namely, at each iteration of the postselected learning procedure, we need not only to decide whether there *exists* an i such that $|\operatorname{Tr}(E_i\rho_t) - \operatorname{Tr}(E_i\rho)|$ is large, but also to *find* such an i if it exists. Fortunately, we can handle this using the "oldest trick in the book" for reducing search problems to decision problems: namely, binary search over the list $E_1, \ldots, E_M$. Doing this correctly requires carefully managing the error budget—as we proceed through binary search, the gap between $\operatorname{Tr}(E_i\rho_t)$ and $\operatorname{Tr}(E_i\rho)$ that we're confident we've found degrades from ε to $\varepsilon - \alpha$ to $\varepsilon - 2\alpha$, etc.—and that's what produces the factor of $\log^4 M$ in the final bound.

1.3 Comparison with Related Work

While we've already discussed a good deal of related work, here we'll compare Theorem 2 directly against some previous results, and explain why those results fall short of what we need. We'll then discuss the recent work of Brandão et al. [14], which builds on this paper to address the computational cost of shadow tomography.

One important inspiration for what we're trying to do, and something we *haven't* yet discussed, is the "Quantum Occam's Razor Theorem," which Aaronson [4] proved in 2006. This result essentially says that quantum states are "learnable" in the PAC (Probably Approximately Correct) sense [29], with respect to any probability distribution over two-outcome measurements, using an amount of sample data that increases only *linearly* with the number of qubits—rather than exponentially, as with traditional quantum state tomography. More formally:

Theorem 3 (Quantum Occam's Razor [4]) *Let ρ be a D-dimensional mixed state, and let μ be any probability distribution or measure over two-outcome measurements. Then given samples $E_1, \ldots, E_M$ drawn independently from μ, with probability at least $1-\delta$, the samples have the following generalization property: any hypothesis state σ such that $|\operatorname{Tr}(E_i\sigma) - \operatorname{Tr}(E_i\rho)| \leq \frac{\gamma\varepsilon}{7}$ for all $i \in [M]$, will also satisfy*

$$\Pr_{E\sim\mu}\left[|\operatorname{Tr}(E\sigma) - \operatorname{Tr}(E\rho)| \leq \varepsilon\right] \geq 1-\gamma,$$

provided we took

$$M \geq \frac{C}{\gamma^2\varepsilon^2}\left(\frac{\log D}{\gamma^2\varepsilon^2}\log^2\frac{1}{\gamma\varepsilon} + \log\frac{1}{\delta}\right)$$

for some large enough constant C.

We could try applying Theorem 3 to the shadow tomography problem. If we do, however, we get only that $\widetilde{O}\left(\frac{\log D}{\gamma^4\varepsilon^4}\right)$ copies of ρ are enough to let us estimate $\operatorname{Tr}(E_i\rho)$ to within error $\pm\varepsilon$, on at least a $1-\gamma$ *fraction* of the measurements $E_1, \ldots, E_M$—rather than on *all* the measurements.

If we want a result that works for all E_i's, then we can instead switch attention to Aaronson's postselected learning theorem [1], the one used to prove the containment $\mathsf{BQP/qpoly} \subseteq \mathsf{PostBQP/poly}$. For completeness, let us restate that theorem in the language of this paper.

Theorem 4 (implicit in [1]; see also [8]) *Let ρ be an unknown D-dimensional mixed state, and let $E_1, \ldots, E_M$ be known two-outcome measurements. Then there exists a classical string, of length $\widetilde{O}\left(\frac{\log D \cdot \log M}{\varepsilon^3}\right)$, from which $\operatorname{Tr}(E_i \rho)$ can be recovered to within additive error $\pm\varepsilon$ for every $i \in [M]$.*

As we mentioned in Section 1.2, Theorem 4 falls short of shadow tomography simply because it's "nondeterministic": it says that a short classical string *exists* from which one could recover the approximate values of every $\operatorname{Tr}(E_i \rho)$, but says nothing about how to find such a string by measuring few copies of ρ.

There's a different way to think about Theorem 4. Along the way to proving the containment $\mathsf{BQP/qpoly} \subseteq \mathsf{QMA/poly}$, Aaronson and Drucker [10] observed the following, by combining a result from classical learning theory with a result from [4] about the "fat-shattering dimension" of quantum states as a hypothesis class.

Theorem 5 (Aaronson and Drucker [10]) *Let $E_1, \ldots, E_M$ be two-outcome measurements on D-dimensional Hilbert space. Then there exists a set S of real functions $f : [M] \to [0,1]$, of cardinality $\left(\frac{M}{\varepsilon}\right)^{O\left(\varepsilon^{-2} \log D\right)}$, such that for every D-dimensional mixed state ρ, there exists an $f \in S$ such that $|f(i) - \operatorname{Tr}(E_i \rho)| \leq \varepsilon$ for all $i \in [M]$.*

Up to a small difference in the parameters, Theorem 5 is equivalent to Theorem 4: either can easily be deduced from the other. The main difference is just that Theorem 4 came with an explicit procedure, based on postselection, for recovering the $\operatorname{Tr}(E_i \rho)$'s from the classical string, whereas Theorem 5 was much less explicit.

It might seem that Theorem 5 would give rise to a shadow tomography procedure, since we'd just need to implement a measurement, say on $O\left(\frac{\log|S|}{\varepsilon^2}\right)$ copies of ρ, that "pulled apart" the different elements of the set S (which is called an *ε-cover*). Unfortunately, we haven't been able to turn this intuition into an algorithm. For while one *can* project a quantum state onto any set of vectors that's sufficiently close to orthogonal—as, for example, in the algorithm of Ettinger, Høyer, and Knill [15] for the hidden subgroup problem—in shadow tomography, there's no guarantee that the state $\rho^{\otimes k}$ being measured *is* close to one of various nearly-orthogonal measurement outcomes, and therefore that it won't be irreparably damaged at an early stage in the measurement process.

Recently, building on the work reported here, Brandão et al. [14] have undertaken an initial investigation of the *computational* complexity of shadow tomography. While we made no attempt to optimize the computational cost of our procedure, a loose estimate is that ours requires performing $\widetilde{O}\left(\frac{M \log D}{\varepsilon^4}\right)$ measurements on copies of ρ. Furthermore, each measurement itself could, in the worst case, require $\Theta\left(D^2\right)$ gates to implement. Our procedure also involves storing and updating a classical description of an amplified hypothesis state, which takes $D^{O\left(\varepsilon^{-2} \log\log D\right)}$ time and space.

By combining our ideas with recent quantum algorithms for semidefinite programming, Brandão et al. [14] have shown how to perform shadow tomography using not only $\operatorname{poly}(\log M, \log D)$ copies of ρ, but also $\widetilde{O}\left(\sqrt{M} L\right) + D^{O(1)}$ quantum gates, where $L = O\left(D^2\right)$ is the maximum length of a circuit to apply a single measurement E_i. This of course improves over our $\widetilde{O}(ML) + D^{O(\log\log D)}$.

If we make some additional assumptions about the measurement matrices E_i—namely, that they have rank at most $\operatorname{polylog} D$; and that for every $i \in [M]$, one can coherently prepare the mixed state $\frac{E_i}{\operatorname{Tr}(E_i)}$, and also compute $\operatorname{Tr}(E_i)$, in time at most $\operatorname{polylog} D$—then Brandão et al. [14] further improve the running time of their algorithm, to $\widetilde{O}\left(\sqrt{M} \operatorname{polylog} D\right)$.

Roughly, Brandão et al. [14] keep much of the structure of our algorithm, except they replace our linear search for informative measurements E_i by Grover-style approximate counting—hence the improvement from $\widetilde{O}(M)$ to $\widetilde{O}\left(\sqrt{M}\right)$. They also replace our postselected learning by the preparation of a Gibbs state, using Jaynes' principle from statistical mechanics. By exploiting recent progress on quantum algorithms for SDPs, Brandão et al. are able to perform the needed manipulations on D-dimensional hypothesis states without ever writing the states explicitly in a classical memory as $D \times D$ matrices, like we do.

In Section 7, we'll discuss the prospects for improving the gate complexity of shadow tomography further, and some possible complexity-theoretic barriers to doing so.

There are many other results in the literature that can be seen, in one way or another, as trying to get around the destructive nature of measurement, or the exponential number of copies needed for state tomography. We won't even attempt a survey here, but briefly, such results often put some additional restriction on the state ρ to be learned: for example, that it's low rank [17], or that it has a succinct classical description of some kind (e.g., that it's a stabilizer state [25]), or that we have an oracle to recognize the state [16]. Of course, shadow tomography requires none of these assumptions.

2 Motivation

Perhaps the most striking way to state Theorem 2 is as follows.

Corollary 6 *Let $|\psi\rangle$ be an unknown n-qubit state, and let p be any fixed polynomial. Then it's possible to estimate $\Pr[C \text{ accepts } |\psi\rangle]$ to within additive error $\pm\varepsilon$, for* ***every*** *quantum circuit C with at most $p(n)$ gates simultaneously, and with $1 - o(1)$ success probability, by a measurement on $(n/\varepsilon)^{O(1)}$ copies of $|\psi\rangle$.*

Here, we're simply combining Theorem 2 with the observation that there are at most $M = (n + p(n))^{O(p(n))}$ different quantum circuits of size at most $p(n)$, assuming a fixed finite gate set without loss of generality.

We've already given some philosophical motivation for this: at bottom we're trying to understand, *to what extent does the destructive nature of quantum measurement force us into an epistemically unsatisfying situation, where we need $\exp(n)$ copies of an n-qubit state $|\psi\rangle$ just to learn $|\psi\rangle$'s basic properties?* Corollary 6 tells us that, as long as the "basic properties" are limited to $|\psi\rangle$'s accept/reject behaviors on quantum circuits of a fixed polynomial size (and to whatever can be deduced from those behaviors), we're *not* in the epistemically unsatisfying situation that might have been feared.

Besides this conceptual point, we hope that Theorem 2 will find experimental applications. In the quest for such applications, it would of course help to tighten the parameters of Theorem 2 (e.g., the exponents in $\frac{\log^4 M}{\varepsilon^5}$); and to find shadow tomography procedures that are less expensive both in computational complexity and in the required measurement apparatus. We'll say more about these issues in Section 7.

In the rest of this section, we'll point out implications of Theorem 2 for several areas of quantum computing theory: quantum money, quantum copy-protected software, and quantum advice and one-way communication. The first of these actually provided the original impetus for this work: as we'll explain, Theorem 2 immediately yields a proof of a basic result called the "tradeoff theorem" for private-key quantum money schemes [6, Section 8.3]. But even where the implications amount to little more than translations of the theorem to other contexts, they illustrate the wide reach of shadow tomography as a concept.

Quantum money. The idea of quantum money—i.e., quantum states that can be traded and verified, but are physically impossible to clone—is one of the oldest ideas in quantum information, having been proposed by Wiesner [30] around 1970. A crucial distinction here is between so-called *public-key* and *private-key* quantum money schemes. See Aaronson and Christiano [9] for formal definitions of these concepts, but briefly: in a public-key money scheme, anyone can efficiently verify a bill $|\$\rangle$ as genuine, whereas in a private-key scheme, verifying a bill requires taking it back to the bank. It's easy to see that, if public-key quantum money is possible at all, then it requires computational assumptions (e.g., that any would-be counterfeiter is limited to polynomial time). While Aaronson and Christiano [9] constructed an oracle relative to which public-key quantum money is possible, it's still unclear whether it's possible in the unrelativized world.

By contrast, in Wiesner's original paper on the subject [30], he proposed a private-key quantum money scheme that was *unconditionally secure* (though a security proof would only be given in 2012, by Molina, Vidick, and Watrous [24]). The central defect of Wiesner's scheme was that it required the bank to maintain a gigantic database, storing a different list of secret measurement bases for every bill in circulation. In 1982, Bennett et al. [13] fixed this defect of Wiesner's scheme, but only by using a pseudorandom function to generate the measurement bases—so that the scheme again required a computational assumption.

In 2009, Aaronson [5] raised the question of whether there's an inherent tradeoff here: that is, does every private-key quantum money scheme require *either* a huge database, or else a computational assumption?[6] He then answered this question in the affirmative (paper still in preparation, but see [6, Section 8.3]). It was while proving this tradeoff theorem that the author was led to formulate the shadow tomography problem.

To see the connection, let's observe an easy corollary of Theorem 2.

Corollary 7 (of Theorem 2) *Consider any private-key quantum money scheme with a single secret key $k \in \{0,1\}^m$ held by the bank; d-qubit bills $|\$\rangle$; and a verification procedure $V(k, |\$\rangle)$ that the bank applies. Then given $\widetilde{O}(dm^4)$ legitimate bills $|\$\rangle$, as well as* $\exp(d,m)$ *computation time, a counterfeiter can estimate* $\Pr[V(k,|\$\rangle)$ *accepts*] *to within additive error $o(1)$, for every $k \in \{0,1\}^m$, with success probability $1-o(1)$.*

Proof. We set $M := 2^m$, and let our list of M two-outcome measurements correspond to $V(k,\cdot)$ for every $k \in \{0,1\}^m$. We set $\rho := |\$\rangle\langle\$|$; this is a D-dimensional state where $D := 2^d$. Then Theorem 2 lets us estimate $\Pr[V(k,|\$\rangle)$ accepts] for every $k \in \{0,1\}^m$ as claimed, using

$$\widetilde{O}\left(\log D \cdot \log^4 M\right) = \widetilde{O}\left(dm^4\right)$$

[6] Actually, he claimed to have an unwritten proof of this, but working out the details took longer than expected, and indeed ultimately relied on the 2016 work of Harrow, Lin, and Montanaro [19].

copies of $|\$\rangle$. ■

We now observe that the tradeoff theorem follows immediately from Corollary 7:

Theorem 8 (Tradeoff Theorem for Quantum Money) *Given any private-key quantum money scheme, with d-qubit bills and an m-bit secret key held by the bank, a counterfeiter can produce additional bills, which pass verification with $1 - o(1)$ probability, given $\widetilde{O}\left(dm^4\right)$ legitimate bills and* $\exp(d, m)$ *computation time. No queries to the bank are needed to produce these bills.*

For given exponential time, the counterfeiter just needs to do a brute-force search (for example, using semidefinite programming) for a state ρ such that

$$|\Pr[V(k, \rho) \text{ accepts}] - \Pr[V(k, |\$\rangle) \text{ accepts}]| = o(1)$$

for every key $k \in \{0,1\}^m$. Such a ρ surely exists, since $|\$\rangle$ itself is one, and given exponential time, the counterfeiter can then prepare ρ as often as it likes. And by assumption, this ρ must be a state that the bank accepts with high probability given the "true" key k^*—*even though the counterfeiter never actually learns k^* itself.*

In [6, Section 8.3], the author took a somewhat different route to proving the tradeoff theorem, simply because he didn't yet possess the shadow tomography theorem. Specifically, he used what in this paper we'll call the "gentle search procedure," and will prove as Lemma 15 along the way to proving Theorem 2. He then combined Lemma 15 with an iterative procedure, which repeatedly cut down the space of "possible keys" k by a constant factor, until averaging over the remaining keys led to a state that the bank accepted with high probability. However, this approach had the drawback that preparing the counterfeit bills required $O(n)$ queries to the bank. Shadow tomography removes that drawback.

Quantum copy-protected software. In 2009, Aaronson [5] introduced the notion of quantum copy-protected software: roughly speaking, an $n^{O(1)}$-qubit quantum state ρ_f that's given to a user, and that lets the user efficiently evaluate a Boolean function $f : \{0,1\}^n \to \{0,1\}$, on any input $x \in \{0,1\}^n$ of the user's choice, but that can't be used to prepare more states with which f can be efficiently evaluated. The analogous classical problem is clearly impossible. But the destructive nature of quantum measurements (or equivalently, the unclonability of quantum states) raises the prospect that, at least with suitable cryptographic assumptions, it could be possible quantumly. And indeed, Aaronson [5] sketched a construction of a quantum oracle U relative to which quantum copy-protection is "generically" possible, meaning that one really *can* have a state $|\psi_f\rangle$ that acts like an unclonable black box for any Boolean function f of one's choice. It remains an outstanding problem to construct *explicit* schemes for quantum copy-protection, which are secure under plausible cryptographic assumptions.

But now suppose that we're interested in quantum programs that simply accept various inputs $x \in \{0,1\}^n$ with specified probabilities $p(x) \in [0,1]$: for example, programs to evaluate partial Boolean functions, or to simulate quantum processes. In that case, we might hope for a copy-protection scheme that was *unconditionally* secure, even against software pirates with unlimited computation time. Furthermore, such a scheme would have the property—possibly desirable to the software vendor!—that the programs would periodically get "used up" even by legitimate use, and need to be replenished. For even if we had $n^{O(1)}$ copies of the program, and used the Gentle Measurement Lemma to estimate the probabilities $p(x)$, we still couldn't always avoid measurements on the "knife edge" between one output behavior and another, which would destroy the copies.

Once again, though, Theorem 2 has the consequence that this gambit fails, so that if quantum copy-protection is possible at all, then it indeed requires computational assumptions.

Corollary 9 (of Theorem 2) *Let ρ be any $n^{O(1)}$-qubit quantum program, which accepts each input $x \in \{0,1\}^n$ with probability $p(x)$. Then given $n^{O(1)}$ copies of ρ and $2^{n^{O(1)}}$ computation time, with $1-o(1)$ success probability we can "pirate" ρ: that is, produce multiple quantum programs, all of which accept input $x \in \{0,1\}^n$ with probability $p(x) \pm o(1)$, and which have the same running time as ρ itself.*

Here we're using the fact that, once we know the approximate acceptance probabilities of ρ on every input $x \in \{0,1\}^n$, in $2^{n^{O(1)}}$ time we can simply use semidefinite programming to brute-force search for an $n^{O(1)}$-qubit state σ that approximates ρ's acceptance probabilities on every x. Indeed, if we further assume that ρ was prepared by a polynomial-size quantum circuit, then in $2^{n^{O(1)}}$ time we can brute-force search for such a circuit as well.

Quantum advice and one-way communication. In 2003, Nishimura and Yamakami [26] defined the complexity class $\mathsf{BQP/qpoly}$, which consists (informally) of all languages that are decidable in bounded-error quantum polynomial time, given a polynomial-size "quantum advice state" $|\psi_n\rangle$ that depends only on the input length n but could otherwise be arbitrary. This is a natural quantum generalization of the classical notion of Karp-Lipton advice, and of the class $\mathsf{P/poly}$. Many results have since been proven about $\mathsf{BQP/qpoly}$ and related classes [1, 3, 10, 11]; and as we discussed in Section 1.2, some of the techniques used to prove those results will also play major roles in this work.

But one basic question remained: given a $\mathsf{BQP/qpoly}$ algorithm, suppose we're given $n^{O(1)}$ copies of the quantum advice state $|\psi_n\rangle$. Can we safely reuse those copies, again and again, for as many inputs $x \in \{0,1\}^n$ as we like? For *deciding a language L*, it's not hard to show that the answer is yes, because of the Gentle Measurement Lemma (Lemma 11 in Section 3). But if we consider *promise problems* (i.e., problems of deciding which of two disjoint sets the input x belongs to, promised that it belongs to one of them), then a new difficulty arises. Namely, what if we use our quantum advice on an input that violates the promise—a possibility that we can't generally avoid if we don't know the promise? Every such use runs the risk of destroying an advice state.

An immediate corollary of Theorem 2 is that we can handle this issue, albeit with a blowup in computation time.

Corollary 10 (of Theorem 2) *Let $\Pi = (\Pi_{\text{YES}}, \Pi_{\text{NO}})$ be a promise problem in $\mathsf{PromiseBQP/qpoly}$. Let A be a quantum algorithm for Π that uses advice states $\{|\psi_n\rangle\}_n$. Then there exists a quantum algorithm, running in $2^{n^{O(1)}}$ time, that uses $|\psi_n\rangle^{\otimes n^{O(1)}}$ as advice, and that approximates $\Pr[A(x, |\psi_n\rangle)$ accepts$]$ to within $\pm o(1)$, for all 2^n inputs $x \in \{0,1\}^n$, with success probability $1-o(1)$. So in particular, this algorithm "generates the complete truth table of Π on inputs of size n," and does so even without being told which inputs satisfy the promise $x \in \Pi_{\text{YES}} \cup \Pi_{\text{NO}}$.*

We can also state Corollary 10 in terms of *quantum one-way communication protocols*. In that case, the corollary says the following. Suppose Alice holds an input $x \in \{0,1\}^n$ and Bob holds an input $y \in \{0,1\}^m$, and they want to compute a partial Boolean function $f : S \to \{0,1\}$, for some

$S \subset \{0,1\}^n \times \{0,1\}^m$. Suppose also that, if Alice sends a q-qubit quantum state $|\psi_x\rangle$ to Bob, then Bob can compute $f(x,y)$ with bounded probability of error, for any $(x,y) \in S$. Then given $\widetilde{O}\left(qm^4\right)$ copies of $|\psi_x\rangle$, Bob can compute $f(x,y)$ for *every* y such that $(x,y) \in S$ simultaneously—again, even though Bob doesn't know which y's satisfy $(x,y) \in S$ (and therefore, which ones might be "dangerous" to measure).

3 Preliminaries

In this section, we collect the (very basic) concepts and results of quantum information that we'll need for this paper. In principle, no quantum information background is needed to read the paper beyond this.

A *mixed state* is the most general kind of state in quantum mechanics, encompassing both superposition and ordinary probabilistic uncertainty. A D-dimensional mixed state ρ is described by a $D \times D$ Hermitian positive semidefinite matrix with $\operatorname{Tr}(\rho) = 1$. If ρ has rank 1, then we call it a *pure state.* At the other extreme, if ρ is diagonal, then it simply describes a classical probability distribution over D outcomes, with $\Pr[i] = \rho_{ii}$. The state $\frac{I}{D}$, corresponding to the uniform distribution over D outcomes, is called the *maximally mixed state.*

Given two mixed states ρ and σ, their *trace distance* is defined as

$$\|\rho - \sigma\|_{\text{tr}} := \frac{1}{2}\sum_i |\lambda_i|,$$

where the λ_i's are the eigenvalues of $\rho - \sigma$. This is a distance metric, which generalizes the variation distance between probability distributions, and which equals the maximum bias with which ρ can be distinguished from σ by a single-shot measurement.

Given a D-dimensional mixed state ρ, one thing we can do is to apply a *two-outcome measurement*, and see whether it accepts or rejects ρ. Such a measurement—technically called a "Positive Operator Valued Measure" or "POVM"—can always be described by a $D \times D$ Hermitian matrix E with all eigenvalues in $[0,1]$ (so in particular, E is positive semidefinite). The measurement E *accepts* ρ with probability $\operatorname{Tr}(E\rho)$, and *rejects* ρ with probability $1 - \operatorname{Tr}(E\rho)$.

The POVM formalism doesn't tell us what happens to ρ after the measurement, and indeed the post-measurement state could in general depend on how E is implemented. However, we have the following extremely useful fact, which was called the "Gentle Measurement Lemma" by Winter [32].

Lemma 11 (Gentle Measurement Lemma [32]) *Let ρ be a mixed state, and let E be a two-outcome measurement such that $\operatorname{Tr}(E\rho) \geq 1 - \varepsilon$. Then after we apply E to ρ, assuming E accepts, we can recover a post-measurement state $\widetilde{\rho}$ such that $\|\widetilde{\rho} - \rho\|_{\text{tr}} \leq 2\sqrt{\varepsilon}$.*

As a historical note, Aaronson [1, 6] proved a variant of Lemma 11, which he called the "Almost As Good As New Lemma"; the main difference is that Aaronson's version doesn't involve conditioning on the case that E accepts.

We'll also need a stronger fact, which goes back at least to Ambainis et al. [12], and which Aaronson [3, 6] called the "Quantum Union Bound." Here we state the strongest version, due to Wilde [31], although a less strong version (involving the bound $M\sqrt{\varepsilon}$ rather than $\sqrt{M\varepsilon}$) would also have worked fine for us.

Lemma 12 (Quantum Union Bound [12, 3, 31]) *Let ρ be a mixed state, and let $E_1, \ldots, E_M$ be two-outcome measurements such that $\operatorname{Tr}(E_i \rho) \geq 1 - \varepsilon$ for all $i \in [M]$. Then if $E_1, \ldots, E_M$ are applied to ρ in succession, the probability that they all accept is at least $1 - 2M\sqrt{\varepsilon}$, and conditioned on all of them accepting, we can recover a post-measurement state $\widetilde{\rho}$ such that*

$$\|\widetilde{\rho} - \rho\|_{\text{tr}} = O\left(\sqrt{M\varepsilon}\right).$$

4 Gentle Search Procedure

We now develop a procedure that takes as input descriptions of two-outcome measurements $E_1, \ldots, E_M$, as well as polylog M copies of an unknown state ρ, and that searches for a measurement E_i that accepts ρ with high probability. In Section 5, we'll then use this procedure as a key subroutine for solving the shadow tomography problem.

Our starting point is a recent result of Harrow, Lin, and Montanaro [19] (their Corollary 11), which we state below for convenience.

Theorem 13 (Harrow, Lin, and Montanaro [19]) *Let ρ be an unknown mixed state, and let $E_1, \ldots, E_M$ be known two-outcome measurements. Suppose we're promised that either*

(i) there exists an $i \in [M]$ such that $\operatorname{Tr}(E_i \rho) \geq 1 - \epsilon$, or else

(ii) $\operatorname{Tr}(E_1 \rho) + \cdots + \operatorname{Tr}(E_M \rho) \leq \Delta M$.

There is a test that uses one copy of ρ, and that accepts with probability at least $(1-\epsilon)^2/7$ in case (i) and with probability at most $4\Delta M$ in case (ii).

Aaronson [3] had previously claimed a version of Theorem 13, which he called the Quantum OR Bound. However, Aaronson's proof had a mistake, which Harrow et al. [19] both identified and fixed.

Briefly, Harrow et al. [19] give two ways to prove Theorem 13. The first way is by adapting the in-place amplification procedure of Marriott and Watrous [23]. The second way is by preparing a control qubit in the state $\frac{|0\rangle+|1\rangle}{\sqrt{2}}$, and then repeatedly applying E_i's to ρ conditional on the control qubit being $|1\rangle$, while also periodically measuring the control qubit in the $\left\{\frac{|0\rangle+|1\rangle}{\sqrt{2}}, \frac{|0\rangle-|1\rangle}{\sqrt{2}}\right\}$ basis to see whether applying the E_i's has decohered the control qubit. Harrow et al. show that, in case (i), *either* some E_i is likely to accept or else the control qubit is likely to be decohered. In case (ii), on the other hand, one can upper-bound the probability that either of these events happen using the Quantum Union Bound (Lemma 12).

Both of Harrow et al.'s procedures perform measurements on ρ that involve an ancilla register, and that are somewhat more complicated than the E_i's themselves. By contrast, the original procedure of Aaronson [3] just applied the E_i's in a random order. It remains an open question whether the simpler procedure is sound.

In any case, by combining Theorem 13 with a small amount of amplification, we can obtain a variant of Theorem 13 that's more directly useful for us, and which we'll call "the" Quantum OR Bound in this paper.

Lemma 14 (Quantum OR Bound) *Let ρ be an unknown mixed state, and let $E_1, \ldots, E_M$ be known two-outcome measurements. Suppose we're promised that either* *(i) there exists an $i \in [M]$ such that* $\operatorname{Tr}(E_i \rho) \geq c$, *or else*

(ii) $\operatorname{Tr}(E_i \rho) \leq c - \varepsilon$ *for all $i \in [M]$.*

We can distinguish these cases, with success probability at least $1 - \delta$, given $\rho^{\otimes k}$ where $k = O\left(\frac{\log 1/\delta}{\varepsilon^2} \log M\right)$.

Proof. This essentially follows by combining Theorem 13 with the Chernoff bound. Assume without loss of generality that $M \geq 50$, and let $\ell = C \frac{\log M}{\varepsilon^2}$ for some sufficiently large constant C. Also, let E_i^* be an amplified measurement that applies E_i to each of ℓ registers, and that accepts if and only if the number of accepting invocations is at least $\left(c - \frac{\varepsilon}{2}\right)\ell$. Then in case (i) we have

$$\operatorname{Tr}\left(E_i^* \rho^{\otimes \ell}\right) \geq 1 - \frac{1}{M^2}$$

for some $i \in [M]$, while in case (ii) we have

$$\operatorname{Tr}\left(E_i^* \rho^{\otimes \ell}\right) \leq \frac{1}{M^2}$$

for all i. So if we apply the procedure of Theorem 13 to $\rho^{\otimes \ell}$, then it accepts with probability at least (say) $\frac{1}{8}$ in case (i), or with probability at most $\frac{4}{M}$ in case (ii).

We now just need $O(\log 1/\delta)$ rounds of further amplification—involving a fresh copy of $\rho^{\otimes \ell}$ in each round—to push these acceptance probabilities to $1 - \delta$ or δ respectively. ■

Note that the procedure of Lemma 14 requires performing collective measurements on $O\left(\frac{\log M}{\varepsilon^2}\right)$ copies of ρ. On the positive side, though, the number of copies has no dependence whatsoever on the Hilbert space dimension D.

Building on Lemma 14, we next want to give a *search* procedure: that is, a procedure that actually *finds* a measurement E_i in our list that accepts ρ with high probability (if there is one), rather than merely telling us whether such an E_i exists. To do this, we'll use the classic trick in computer science for reducing search problems to decision problems: namely, binary search over the list $E_1, \ldots, E_M$.

The subtlety is that, as we run binary search, our lower bound on the acceptance probability of the measurement E_i that we're isolating degrades at each level of the recursion, while the error probability builds up. Also, we need fresh copies of ρ at each level of the recursion. Handling these issues will yield a procedure that uses roughly $\frac{\log^4 M}{\varepsilon^2}$ copies of ρ, which we suspect is not tight.

Lemma 15 (Gentle Search) *Let ρ be an unknown mixed state, and let $E_1, \ldots, E_M$ be known two-outcome measurements. Suppose there exists an $i \in [M]$ such that* $\operatorname{Tr}(E_i \rho) \geq c$. *Then we can find a $j \in [M]$ such that* $\operatorname{Tr}(E_j \rho) \geq c - \varepsilon$, *with success probability at least $1 - \delta$, given $\rho^{\otimes k}$ where*

$$k = O\left(\frac{\log^4 M}{\varepsilon^2}\left(\log \log M + \log \frac{1}{\delta}\right)\right).$$

Proof. Assume without loss of generality that M is a power of 2. We will apply Lemma 14 recursively, using binary search to zero in on a j such that $\operatorname{Tr}(E_j \rho) \geq c - \varepsilon$.

Divide the measurements into two sets, $S_1 = \left\{E_1, \ldots, E_{M/2}\right\}$ and $S_2 = \left\{E_{M/2+1}, \ldots, E_M\right\}$. Also, let $\alpha := \frac{\varepsilon}{\log_2 M}$, and let $\beta := \frac{\delta}{\log_2 M}$. Then as a first step, we call the subroutine from Lemma 14 to check, with success probability at least $1 - \beta$, whether

(i) there exists an $E \in S_1$ such that $\mathrm{Tr}\,(E\rho) \geq c$ or

(ii) $\mathrm{Tr}\,(E\rho) \leq c - \alpha$ for all $E \in S_1$,

promised that one of these is the case.

Note that the promise could be violated—but this simply means that, if the subroutine returns (i), then we can assume only that there exists an $E \in S_1$ such that $\mathrm{Tr}\,(E\rho) \geq c - \alpha$.

Thus, if the subroutine returns (i), then we recurse on S_1. That is, we divide S_1 into two sets both of size $\frac{M}{4}$, and then use Lemma 14 to find (again with success probability at least $1 - \beta$) a set that contains an E such that $\mathrm{Tr}\,(E\rho) \geq c - 2\alpha$, assuming now that one of the two sets contains an E such that $\mathrm{Tr}\,(E\rho) \geq c - \alpha$. If the subroutine returns (ii), then we do the same but with S_2.

We continue recursing in this way, identifying a set of size $\frac{M}{8}$ that contains an E such that $\mathrm{Tr}\,(E\rho) \geq c - 3\alpha$, then a set of size $\frac{M}{16}$ that contains an E such that $\mathrm{Tr}\,(E\rho) \geq c - 4\alpha$, and so on, until we reach a singleton set. This gives us our index j such that

$$\begin{aligned}\mathrm{Tr}\,(E_j\rho) &\geq c - \alpha \log_2 M \\ &= c - \varepsilon.\end{aligned}$$

By the union bound, together with the promise that there exists an $i \in [M]$ such that $\mathrm{Tr}\,(E_i\rho) \geq c$, the whole procedure succeeds with probability at least $1 - \beta \log_2 M = 1 - \delta$. Meanwhile, within each of the $\log_2 M$ iterations of this procedure, Lemma 14 tells us that the number of copies of ρ that we need is

$$\begin{aligned}O\left(\frac{\log 1/\beta}{\alpha^2} \log M\right) &= O\left(\frac{\log \frac{\log M}{\delta}}{(\varepsilon/\log M)^2} \log M\right) \\ &= O\left(\frac{\log^3 M}{\varepsilon^2}\left(\log\log M + \log \frac{1}{\delta}\right)\right).\end{aligned}$$

Therefore the total number of copies needed is

$$O\left(\frac{\log^4 M}{\varepsilon^2}\left(\log\log M + \log \frac{1}{\delta}\right)\right).$$

■

5 Main Result

We're now ready to prove Theorem 2, which we restate for convenience. Given an unknown D-dimensional mixed state ρ, and known two-outcome measurements $E_1, \ldots, E_M$, we'll show how to approximate $\mathrm{Tr}\,(E_i\rho)$ to within additive error $\pm\varepsilon$, with success probability at least $1 - \delta$, given $\rho^{\otimes k}$ where

$$k = \widetilde{O}\left(\frac{\log 1/\delta}{\varepsilon^5} \log^4 M \cdot \log D\right).$$

Afterwards we'll remark on how to improve the ε^{-5} to ε^{-4}, using a recent algorithm of Aaronson et al. [7] for online learning of quantum states.

Proof of Theorem 2. At a high level, we'll use an iterative procedure similar to the multiplicative weights update method: one that, at the t^{th} iteration, maintains a current hypothesis ρ_t about ρ.

Since we're not concerned here with computation time, we can assume that the entire $D \times D$ density matrix of ρ_t is stored to suitable precision in a classical memory, so that we can perform updates (in particular, involving postselection) that wouldn't be possible were ρ_t an actual physical state.

Our initial hypothesis is that ρ is just the maximally mixed state, $\rho_0 = \frac{I}{D}$. Of course this hypothesis is unlikely to give adequate predictions—but by using Lemma 15 as a subroutine, we'll repeatedly refine the current hypothesis, ρ_t, to a "better" hypothesis ρ_{t+1}. The procedure will terminate when we reach a hypothesis ρ_T such that

$$|\operatorname{Tr}(E_i \rho_T) - \operatorname{Tr}(E_i \rho)| \leq \varepsilon$$

for all $i \in [M]$. For at that point, for each i, we can just output $\operatorname{Tr}(E_i \rho_T)$ as our additive estimate for $\operatorname{Tr}(E_i \rho)$.

At each iteration t, we'll use "fresh" copies of ρ, in the course of refining ρ_t to ρ_{t+1}. Thus, we'll need to upper-bound both the total number T of iterations until termination, *and* the number of copies of ρ used in a given iteration.

A key technical ingredient will be amplification. Let

$$q := \frac{C}{\varepsilon^2} \left(\log \log D + \log \frac{1}{\varepsilon} \right)$$

for some suitable constant C, and let $\rho^* := \rho^{\otimes q}$. Then, strictly speaking, our procedure will maintain a hypothesis ρ_t^* about ρ^*: the initial hypothesis is the maximally mixed state $\rho_0^* = \frac{I}{D^q}$; then we'll refine the hypothesis to ρ_1^*, ρ_2^*, and so on. At any point, we let ρ_t be the D-dimensional state obtained by choosing a register of ρ_t^* uniformly at random, and tracing out the remaining $q-1$ registers.

Given the hypothesis ρ_t, for each $i \in [M]$, let $E_{i,t,+}^*$ be a two-outcome measurement on $\rho^{\otimes q}$ that applies E_i to each of the q registers, and that accepts if and only if the number of accepting invocations is at least $\left(\operatorname{Tr}(E_i \rho_t) + \frac{3\varepsilon}{4}\right) q$. Likewise, let the measurement $E_{i,t,-}^*$ apply E_i to each of the q registers, and accept if and only if the number of accepting invocations is at most $\left(\operatorname{Tr}(E_i \rho_t) - \frac{3\varepsilon}{4}\right) q$.

Suppose $\operatorname{Tr}(E_i \rho) \geq \operatorname{Tr}(E_i \rho_t) + \varepsilon$. Then by a Chernoff bound, we certainly have $\operatorname{Tr}\left(E_{i,t,+}^* \rho^*\right) \geq \frac{5}{6}$, provided the constant C was sufficiently large: indeed, for this we need only that q grows at least like $\frac{C}{\varepsilon^2}$. Likewise, if $\operatorname{Tr}(E_i \rho) \leq \operatorname{Tr}(E_i \rho_t) - \varepsilon$, then $\operatorname{Tr}\left(E_{i,t,-}^* \rho^*\right) \geq \frac{5}{6}$.

On the other hand, suppose $|\operatorname{Tr}(E_i \rho) - \operatorname{Tr}(E_i \rho_t)| \leq \frac{\varepsilon}{2}$. Then again by a Chernoff bound, we have $\operatorname{Tr}\left(E_{i,t,+}^* \rho^*\right) \leq \frac{1}{3}$ and $\operatorname{Tr}\left(E_{i,t,-}^* \rho^*\right) \leq \frac{1}{3}$, provided the constant C is sufficiently large.

We can now give the procedure to update the hypothesis ρ_t^* to ρ_{t+1}^*. Let $\beta := \frac{\delta \varepsilon^4}{\log^2 D}$. Then at each iteration $t \geq 0$, we do the following:

- Use Lemma 15 to search for an index $j \in [M]$ such that $\operatorname{Tr}\left(E_{j,t,+}^* \rho^*\right) \geq \frac{2}{3}$, promised that there exists such a j with $\operatorname{Tr}\left(E_{j,t,+}^* \rho^*\right) \geq \frac{5}{6}$ (which we call the $+$ case); or for a $j \in [M]$ such that $\operatorname{Tr}\left(E_{j,t,-}^* \rho^*\right) \geq \frac{2}{3}$, promised that there exists a j such that $\operatorname{Tr}\left(E_{j,t,-}^* \rho^*\right) \geq \frac{5}{6}$ (which we call the $-$ case). Set the parameters so that, assuming that one or both promises hold, the search succeeds with probability at least $1 - \beta$.

- If no j is found that satisfies either condition, then halt and return ρ_t as the hypothesis state: in other words, return $\operatorname{Tr}(E_i \rho_t)$ as the estimate for $\operatorname{Tr}(E_i \rho)$, for all $i \in [M]$.

- Otherwise, if a j *is* found satisfying one of the conditions, then let F_t be a measurement that applies E_j to each of the q registers, and that accepts if and only if the number of accepting invocations is at least $\left(\operatorname{Tr}(E_j \rho_t) + \frac{\varepsilon}{4}\right) q$ (in the $+$ case), or at most $\left(\operatorname{Tr}(E_j \rho_t) - \frac{\varepsilon}{4}\right) q$ (in the $-$ case).

- Let ρ_{t+1}^* be the state obtained by starting from ρ_t^*, and then postselecting on F_t accepting.

Our central task is to prove an upper bound, T, on the number of iterations of the above procedure until it terminates with states ρ_T^* and ρ_T such that $|\operatorname{Tr}(E_i \rho_T) - \operatorname{Tr}(E_i \rho)| \leq \varepsilon$ for all $i \in [M]$.

Assume, in what follows, that every invocation of Lemma 15 succeeds in finding a j such that $\operatorname{Tr}\left(E_{j,t,+}^* \rho^*\right) \geq \frac{2}{3}$ or $\operatorname{Tr}\left(E_{j,t,-}^* \rho^*\right) \geq \frac{2}{3}$. Later we will lower-bound the probability that this indeed happens.

To upper-bound T, let

$$p_t = \operatorname{Tr}(F_0 \rho_0^*) \cdot \cdots \cdot \operatorname{Tr}\left(F_{t-1} \rho_{t-1}^*\right)$$

be the probability that the first t postselection steps all succeed.

Then, on the one hand, we claim that $p_{t+1} \leq (1 - \Omega(\varepsilon)) p_t$ for all t. To see this, note that when we run F_t on the state ρ_t^*, the expected number of invocations of E_j that accept is exactly $\operatorname{Tr}(E_j \rho_t) q$. We can't treat these invocations as independent events, because the state ρ_t^* could be correlated or entangled across its q registers in some unknown way. Regardless of correlations, though, by Markov's inequality, in the $+$ case we have

$$\frac{p_{t+1}}{p_t} = \operatorname{Tr}(F_t \rho_t^*) \leq \frac{\operatorname{Tr}(E_j \rho_t) q}{\left(\operatorname{Tr}(E_j \rho_t) + \frac{\varepsilon}{4}\right) q} = 1 - \Omega(\varepsilon).$$

Similarly, in the $-$ case we have

$$\frac{p_{t+1}}{p_t} = \operatorname{Tr}(F_t \rho_t^*) \leq \frac{(1 - \operatorname{Tr}(E_j \rho_t)) q}{\left(1 - \operatorname{Tr}(E_j \rho_t) + \frac{\varepsilon}{4}\right) q} = 1 - \Omega(\varepsilon).$$

Thus $p_t \leq (1 - \varepsilon)^{\Omega(t)}$.

On the other hand, we also claim that $p_t \geq \frac{0.9}{D^q}$ for all $t = o\left(\frac{\log^2 D}{\varepsilon^4}\right)$. To see this: suppose that at iteration t, we had used ρ^* rather than ρ_t^* as the hypothesis state—except still choosing the index $j \in [M]$ as if the hypothesis was ρ_t^*. In that case, when we applied F_t to ρ^*, the expected number of accepting invocations of E_j would be exactly $\operatorname{Tr}(E_j \rho) q$.

Consider for concreteness the $+$ case; the $-$ case is precisely analogous. By the assumption that the search for j succeeded, we have $\operatorname{Tr}\left(E_{j,t,+}^* \rho^*\right) \geq \frac{2}{3}$. In other words: when we apply E_j to q copies of ρ, the number of invocations that accept is at least $\left(\operatorname{Tr}(E_i \rho_t) + \frac{3\varepsilon}{4}\right) q$, with probability at least $\frac{2}{3}$. Recall that $q \geq \frac{C}{\varepsilon^2}$ for some sufficiently large constant C. So since the q copies of ρ really *are* independent, in this case we can use a Chernoff bound to conclude that $\operatorname{Tr}(E_i \rho) > \operatorname{Tr}(E_i \rho_t) + \frac{\varepsilon}{2}$.

Now we consider $1 - \operatorname{Tr}(F_t \rho^*)$: that is, the probability that F_t rejects ρ^*. This is just the probability that, when we apply E_j to q copies of ρ, the number of invocations that accept is less than $\left(\operatorname{Tr}(E_j \rho_t) + \frac{\varepsilon}{4}\right) q$. Since the copies of ρ are independent, and since (by the above) the expected number of accepting invocations is at least $\left(\operatorname{Tr}\left(E_i\rho_t\right)+\frac{\varepsilon}{2}\right)q$, we can again use a Chernoff bound to conclude that

$$\begin{aligned}1-\operatorname{Tr}\left(F_t\rho^*\right) &\leq \exp\left(-\Omega\left(\varepsilon^2 q\right)\right)\\ &\leq \exp\left(-\Omega\left(\varepsilon^2\cdot\frac{C}{\varepsilon^2}\left(\log\log D+\log\frac{1}{\varepsilon}\right)\right)\right)\\ &\leq \frac{\varepsilon^4}{\log^2 D},\end{aligned}$$

provided we took the constant C sufficiently large.

By Lemma 12 (the Quantum Union Bound), this means that, even if we applied $F_0,\ldots,F_{T-1}$ to ρ^* in succession, the probability that *any* of them would reject is at most

$$O\left(\sqrt{\frac{T\varepsilon^4}{\log^2 D}}\right)=O\left(\frac{\sqrt{T}\varepsilon^2}{\log D}\right).$$

Hence, so as long as $T=o\left(\frac{\log^2 D}{\varepsilon^4}\right)$, all of the F_t's accept ρ^* with probability at least (say) 0.9.

But we can always decompose the maximally mixed state, $\rho_0^*=\frac{I}{D^q}$, as

$$\rho_0^*=\frac{1}{D^q}\rho^*+\left(1-\frac{1}{D^q}\right)\sigma,$$

where σ is some mixed state. This means that, in the "real" situation (i.e., when we run the procedure with initial state ρ_0^*), all of the F_t's accept ρ_0^* with probability at least $\frac{0.9}{D^q}$, as claimed.

Combining the two claims above—namely, $p_t\leq(1-\varepsilon)^{\Omega(t)}$ and $p_t\geq\frac{0.9}{D^q}$—we get

$$(1-\varepsilon)^{\Omega(t)}\geq\frac{0.9}{D^q}.$$

Solving for t now yields

$$\begin{aligned}t&=O\left(\frac{q\log D}{\varepsilon}\right)\\ &=O\left(\frac{\log D}{\varepsilon}\cdot\frac{C}{\varepsilon^2}\left(\log\log D+\log\frac{1}{\varepsilon}\right)\right).\end{aligned}$$

This then gives us the desired upper bound on T, and justifies the assumption we made before that $T=o\left(\frac{\log^2 D}{\varepsilon^4}\right)$.

Meanwhile, by Lemma 15 together with the union bound, the probability that all T invocations of Lemma 15 succeed at finding a suitable index j is at least

$$1-T\beta=1-o\left(\frac{\log^2 D}{\varepsilon^4}\cdot\frac{\delta\varepsilon^4}{\log^2 D}\right)\geq 1-\delta,$$

as needed.

Finally, we upper-bound the total number of copies of ρ used by the procedure. Within each iteration t, the bound of Lemma 15 tells us that we need

$$\ell=O\left(\log^4 M\left(\log\log M+\log\frac{1}{\beta}\right)\right)=O\left(\log^4 M\left(\log\log M+\log\log D+\log\frac{1}{\varepsilon}+\log\frac{1}{\delta}\right)\right)$$

copies of the amplified state ρ^*. Since $\rho^* = \rho^{\otimes q}$, this translates to $q\ell$ copies of ρ itself in each of the T iterations, or

$$\begin{aligned}
Tq\ell &= O\left(\frac{q \log D}{\varepsilon} \cdot q\ell\right) \\
&= O\left(\frac{\log D}{\varepsilon} \cdot q^2 \cdot \ell\right) \\
&= O\left(\frac{\log D}{\varepsilon} \cdot \left(\frac{\log\log D + \log \frac{1}{\varepsilon}}{\varepsilon^2}\right)^2 \cdot \log^4 M \left(\log\log M + \log\log D + \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right) \\
&= \widetilde{O}\left(\frac{\log 1/\delta}{\varepsilon^5} \cdot \log^4 M \cdot \log D\right)
\end{aligned}$$

copies of ρ total. ■

Let us briefly indicate how recent work by Aaronson et al. [7] can be used to improve the dependence on ε in Theorem 2 from $1/\varepsilon^5$ to $1/\varepsilon^4$. Apart from some low-order terms needed to amplify success probabilities, the proof of Theorem 2 can be seen as simply a composition of two algorithms:

(1) an algorithm for learning a D-dimensional mixed state ρ, starting with the initial hypothesis $\rho_0 = \frac{I}{D}$ and then repeatedly refining it, given a sequence of two-outcome measurements on which the current hypothesis ρ_t is wrong by more then ε,

(2) an algorithm (namely, the Gentle Search Procedure of Lemma 15) for actually *finding* the measurements on which the current hypothesis is wrong by more than ε, in the context of shadow tomography.

Our algorithm (1) uses $\widetilde{O}\left(\frac{\log D}{\varepsilon^3}\right)$ refinement iterations, while algorithm (2)—which gets invoked inside every iteration—uses $\widetilde{O}\left(\frac{\log^4 M}{\varepsilon^2}\right)$ copies of ρ. Multiplying these two bounds is what produces our final sample complexity of

$$\widetilde{O}\left(\frac{\log^4 M \cdot \log D}{\varepsilon^5}\right).$$

Motivated by a slightly different setting (namely, online learning of quantum states), Aaronson et al. [8] have now given a black-box way, using matrix multiplicative weights and convex optimization, to improve the number of iterations in (1) to $\widetilde{O}\left(\frac{\log D}{\varepsilon^2}\right)$, which matches an information-theoretic lower bound for (1) (see [4]). Using that in place of the postselection procedure of Theorem 2, and combining with the Gentle Search Procedure, yields a slightly improved bound of

$$\widetilde{O}\left(\frac{\log^4 M \cdot \log D}{\varepsilon^4}\right)$$

on the sample complexity of shadow tomography.

6 Lower Bound

We now show that, for purely information-theoretic reasons, any solution to the shadow tomography problem requires at least $\Omega\left(\frac{\min\{D^2,\log M\}}{\varepsilon^2}\right)$ copies of ρ. Indeed, even the classical special case of the problem requires $\Omega\left(\frac{\min\{D,\log M\}}{\varepsilon^2}\right)$ copies. These are the best lower bounds for shadow tomography that we currently know.

6.1 Classical Special Case

We start with the special case where we're given k samples from an unknown distribution $\mathcal{D}$ over a D-element set, and our goal is to learn $\operatorname{E}_{x\sim\mathcal{D}}[f_i(x)]$ to within additive error $\pm\varepsilon$, for each of M known Boolean functions $f_1,\ldots,f_M:[D]\rightarrow\{0,1\}$.

Theorem 16 *Any strategy for shadow tomography—i.e., for estimating* $\operatorname{Tr}(E_i\rho)$ *to within additive error* ε *for all* $i\in[M]$*, with success probability at least (say)* $2/3$*—requires* $\Omega\left(\frac{\min\{D,\log M\}}{\varepsilon^2}\right)$ *copies of the* D*-dimensional mixed state* ρ*. Furthermore, this is true even for the classical special case (i.e., where* ρ *and the* E_i*'s are all diagonal).*

Proof. Set $N:=\lfloor\min\{D,\log_2 M\}\rfloor$. Our distributions will be over N-element sets. Also, for some constant $c\in(1,2)$, set $K:=\lfloor c^N\rfloor$, so that $K\leq M$. We will have K two-outcome measurements.

The first step is to choose K subsets $S_1,\ldots,S_K\subset[N]$ uniformly and independently, subject to the constraint that $|S_i|=\frac{N}{2}$ for each $i\in[K]$. As long as c is sufficiently small, by a Chernoff bound and union bound, it's not hard to see that with probability $1-o(1)$ over the choice of S_i's, we'll have

$$\left||S_i\cap S_j|-\frac{N}{4}\right|\leq\frac{N}{12} \tag{1}$$

for all $i\neq j$. So fix a choice of subsets S_i for which this happens.

Next, for each $i\in[K]$, let $\mathcal{D}_i$ be the probability distribution over $x\in[N]$ defined as follows: choose x uniformly from S_i with probability $\frac{1}{2}+3\varepsilon$, or uniformly from $[N]\setminus S_i$ with probability $\frac{1}{2}-3\varepsilon$. Also, for each $i\in[K]$, let E_i be the standard basis measurement that accepts each basis state $x\in[N]$ if $x\in S_i$ and rejects it otherwise. Then by construction, for all $i\in[K]$, we have

$$\Pr_{x\sim\mathcal{D}_i}[E_i(x)\text{ accepts}]=\frac{1}{2}+3\varepsilon.$$

Also, by (1), for all $i\neq j$ we have

$$\left|\Pr_{x\sim\mathcal{D}_i}[E_j(x)\text{ accepts}]-\frac{1}{2}\right|\leq\frac{\varepsilon}{2}.$$

It follows that, if we can estimate $\Pr_{x\sim\mathcal{D}_i}[E_j(x)\text{ accepts}]$ to within additive error $\pm\varepsilon$ for every $j\in[K]$, that is enough to determine $i\in[K]$.

Notice that, if we choose $i\in[K]$ uniformly at random, then it contains $\log_2(K)=\Omega(N)$ bits of information. Thus, let $X=(x_1,\ldots,x_T)$ consist of T independent samples from $\mathcal{D}_i$. Then in order for it to be information-theoretically *possible* to learn i from X, the mutual information $\mathrm{I}(X;i)$ must be at least $\log_2(K)$. We can now write

$$\begin{aligned}
\mathrm{I}(X;i) &= \mathrm{H}(X) - \mathrm{H}(X|i) \\
&= \sum_{t=1}^{T} \left(\mathrm{H}(x_t \mid x_1,\ldots,x_{t-1}) - \mathrm{H}(x_t \mid i, x_1,\ldots,x_{t-1}) \right) \\
&= \sum_{t=1}^{T} \left(\mathrm{H}(x_t \mid x_1,\ldots,x_{t-1}) - \mathrm{H}(x_t \mid i) \right) \\
&\leq \sum_{t=1}^{T} \left(\mathrm{H}(x_t) - \mathrm{H}(x_t \mid i) \right) \\
&\leq \sum_{t=1}^{T} \left(\log_2 N - \mathrm{H}(\mathcal{D}_i) \right).
\end{aligned}$$

Here the third line follows since x_t has no further dependence on $x_1,\ldots,x_{t-1}$ once we've already conditioned on i, and the last since a distribution over $[N]$ can have entropy at most $\log_2 N$.

Now for all $i \in [K]$,

$$\begin{aligned}
\mathrm{H}(\mathcal{D}_i) &= \sum_{x=1}^{N} \Pr_{\mathcal{D}_i}[x] \log_2 \frac{1}{\Pr_{\mathcal{D}_i}[x]} \\
&= \frac{N}{2}\left(\frac{1/2+3\varepsilon}{N/2}\right)\log_2\left(\frac{N/2}{1/2+3\varepsilon}\right) + \frac{N}{2}\left(\frac{1/2-3\varepsilon}{N/2}\right)\log_2\left(\frac{N/2}{1/2-3\varepsilon}\right) \\
&= \log_2 N - \left[1 - \left(\frac{1}{2}+3\varepsilon\right)\log_2\left(\frac{1}{1/2+3\varepsilon}\right) - \left(\frac{1}{2}-3\varepsilon\right)\log_2\left(\frac{1}{1/2-3\varepsilon}\right)\right] \\
&\geq \log_2 N - O\left(\varepsilon^2\right).
\end{aligned}$$

Combining,

$$\mathrm{I}(X;i) = O\left(T\varepsilon^2\right).$$

We conclude that, in order to achieve mutual information $\mathrm{I}(X;i) = \Omega(N)$, so that X can determine i,

$$T = \Omega\left(\frac{N}{\varepsilon^2}\right) = \Omega\left(\frac{\min\{D, \log M\}}{\varepsilon^2}\right)$$

samples from $\mathcal{D}_i$ are information-theoretically necessary. ■

Let's observe that, in the classical special case, Theorem 16 has a matching upper bound.

Proposition 17 *Let $\mathcal{D}$ be an unknown distribution over $[D]$, and let $f_1,\ldots,f_M : [D] \to \{0,1\}$ be known Boolean functions. Then given*

$$O\left(\frac{1}{\varepsilon^2}\min\left\{D\log\frac{1}{\delta}, \log\frac{M}{\delta}\right\}\right)$$

independent samples from $\mathcal{D}$, it's possible to estimate $\mathrm{E}_{x\sim\mathcal{D}}[f_i(x)]$ to within additive error $\pm\varepsilon$ for each $i \in [M]$, with success probability at least $1-\delta$.

Proof. First, to achieve $O\left(\frac{D\log 1/\delta}{\varepsilon^2}\right)$: it's a folklore fact (see for example [22]) that, given an unknown distribution $\mathcal{D}$ over $[D]$, with high probability, $O\left(\frac{D}{\varepsilon^2}\right)$ samples suffice to learn $\mathcal{D}$ up to ε error in variation distance. This probability can then be amplified to $1-\delta$ at the cost of $O\left(\log\frac{1}{\delta}\right)$ repetitions. And of course, once $\mathcal{D}$ has been so learned, we can then estimate any $\mathrm{E}_{x\sim\mathcal{D}}\left[f_i\left(x\right)\right]$ we like to within additive error $\pm\varepsilon$.

Second, to achieve $O\left(\frac{\log M/\delta}{\varepsilon^2}\right)$: the strategy is simply, for each $i\in[M]$, to output the empirical mean of $f_i\left(x\right)$ on the observed samples. By a Chernoff bound, this strategy fails for any particular i with probability at most

$$\exp\left(-\varepsilon^2\frac{\log M/\delta}{\varepsilon^2}\right)\leq\frac{\delta}{M},$$

provided the constant in the big-O is sufficiently large. So by the union bound, it succeeds for every i simultaneously with probability at least $1-\delta$. ■

We now derive some additional consequences from the proof of Theorem 16. Note that, because recovering the secret index i boils down to identifying a *single* measurement E_j such that

$$\Pr_{x\sim\mathcal{D}_i}\left[E_j\left(x\right)\text{ accepts}\right]>\frac{1}{2}+\varepsilon,$$

the example from the proof also shows that the search problem of Lemma 15, considered by itself, already requires $\Omega\left(\frac{\min\{D,\log M\}}{\varepsilon^2}\right)$ copies of ρ—again, even in the classical special case, where ρ and the E_j's are diagonal.

Indeed, we now make a further observation: for the specific example in Theorem 16, going from Lemma 14 (the Quantum OR Bound) to Lemma 15 (the gentle search procedure) requires almost no blowup in the number of copies of ρ. This is true for two reasons. First, ρ is classical, so in the proof of Lemma 15, we can reuse the same copies of ρ from one binary search iteration to the next. Second, the example assumed a "promise gap"—i.e., for all i,j, the distribution $\mathcal{D}_i$ is either 3ε-biased toward E_j accepting or else not even ε-biased toward it—so there is no need to replace ε by $\frac{\varepsilon}{\log M}$. The only amplification we need is $\log\log M$ repetitions, to push the failure probability per binary search iteration down to $\frac{1}{\log M}$.

We conclude that the Quantum OR Bound—again, even in the classical special case—requires $\Omega\left(\frac{\min\{D,\log M\}}{\varepsilon^2\log\log M}\right)$ copies of ρ, since otherwise we could solve the search problem on the Hadamard example with $o\left(\frac{\min\{D,\log M\}}{\varepsilon^2}\right)$ samples, contradicting our previous reasoning. In other words, Lemma 14 is close to tight.

6.2 General Case

We now show that, when the goal is to learn D-dimensional quantum mixed states, Theorem 16 can be tightened to $\Omega\left(\frac{\min\left\{D^2,\log M\right\}}{\varepsilon^2}\right)$. Note that this subsumes the lower bounds of O'Donnell and Wright [27] and Haah et al. [18], reducing to the latter as we let M be arbitrarily large.

We'll need the following lemma, which is a special case of a more general result proved by Hayden, Leung, and Winter [20].

Lemma 18 ([20, Lemma III.5]) *Let N be even, and let S be a subspace of $\mathbb{C}^N$ chosen uniformly at random subject to $\dim\left(S\right)=\frac{N}{2}$. Let $\mathbb{P}_S$ be the projection onto S, and $\rho_S=\frac{2}{N}\mathbb{P}_S$ be the maximally* *mixed state projected onto S. Then for any fixed subspace $T \leq \mathbb{C}^N$ of dimension $\frac{N}{2}$—for example, the one spanned by $|1\rangle, \ldots, |N/2\rangle$—we have*

$$\Pr_S \left[\left| \operatorname{Tr}(\mathbb{P}_T \rho_S) - \frac{1}{4} \right| > \frac{1}{20} \right] \leq e^{-\Omega(N^2)}.$$

We can now prove the lower bound.

Theorem 19 *Any strategy for shadow tomography—i.e., for estimating $\operatorname{Tr}(E_i \rho)$ to within additive error ε for all $i \in [M]$, with success probability at least (say) $2/3$—requires $\Omega\left(\frac{\min\{D^2, \log M\}}{\varepsilon^2}\right)$ copies of the D-dimensional mixed state ρ.*

Proof. Set $N := \left\lfloor \min\left\{D, \sqrt{\log_2 M}\right\} \right\rfloor$. Our mixed states will be N-dimensional. Also, for some constant $c \in (1,2)$, set $K := \left\lfloor c^{N^2} \right\rfloor$, so that $K \leq M$. We will have K two-outcome measurements.

The first step is to choose K subspaces $S_1, \ldots, S_K \leq \mathbb{C}^N$ Haar-randomly and independently, subject to the constraint that $\dim(S_i) = \frac{N}{2}$ for each $i \in [K]$. Let $\mathbb{P}_i$ be the projection onto S_i, and let $\rho_i := \frac{2}{N}\mathbb{P}_i$ be the maximally mixed state projected onto S_i. Then as long as c is sufficiently small, Lemma 18 tells us that with probability $1 - o(1)$ over the choice of S_i's, we'll have

$$\left| \operatorname{Tr}(\mathbb{P}_i \rho_j) - \frac{1}{2} \right| \leq \frac{1}{12} \tag{2}$$

for all $i \neq j$. So fix a choice of subspaces S_i for which this happens.

Next, for each $i \in [K]$, define the mixed state

$$\sigma_i := (1 - 6\varepsilon)\frac{\mathbb{I}}{N} + 6\varepsilon \rho_i.$$

Our two-outcome POVMs will simply be the projectors $\mathbb{P}_1, \ldots, \mathbb{P}_K$. Then by construction, for all $i \in [K]$, we have

$$\begin{aligned}
\operatorname{Tr}(\mathbb{P}_i \sigma_i) &= (1 - 6\varepsilon)\frac{\operatorname{Tr}(\mathbb{P}_i)}{N} + 6\varepsilon \operatorname{Tr}(\mathbb{P}_i \rho_i) \\
&= (1 - 6\varepsilon)\frac{N/2}{N} + 6\varepsilon \\
&= \frac{1}{2} + 3\varepsilon.
\end{aligned}$$

Also, by (2), for all $i \neq j$ we have

$$\begin{aligned}
\left| \operatorname{Tr}(\mathbb{P}_j \sigma_i) - \frac{1}{2} \right| &= \left| \left((1 - 6\varepsilon)\frac{\operatorname{Tr}(\mathbb{P}_j)}{N} - \left(\frac{1}{2} - 3\varepsilon\right) \right) + \left(6\varepsilon \operatorname{Tr}(\mathbb{P}_j \rho_i) - 3\varepsilon\right) \right| \\
&= 6\varepsilon \left| \operatorname{Tr}(\mathbb{P}_j \rho_i) - \frac{1}{2} \right| \\
&\leq \frac{\varepsilon}{2}.
\end{aligned}$$

It follows that, if we can estimate $\operatorname{Tr}(\mathbb{P}_j \sigma_i)$ to within additive error $\pm\varepsilon$ for every $j \in [K]$, that is enough to determine $i \in [K]$.

Notice that, if we choose $i \in [K]$ uniformly at random, then it contains $\log_2(K) = \Omega\left(N^2\right)$ bits of information. Thus, let

$$\zeta := \mathrm{E}_{i \in [K]}\left[\sigma_i^{\otimes T}\right].$$

Then in order for it to be information-theoretically *possible* to learn i from ζ, the quantum mutual information $\mathrm{I}(\zeta; i)$ must be at least $\log_2(K)$. Since i is classical, we can now write

$$\begin{aligned}
\mathrm{I}(\zeta; i) &= \mathrm{S}(\zeta) - \mathrm{S}(\zeta \mid i) \\
&= \mathrm{S}(\zeta) - \mathrm{S}\left(\sigma_i^{\otimes T}\right) \\
&\leq T\left(\log_2 N - \mathrm{S}(\sigma_i)\right),
\end{aligned}$$

where S is von Neumann entropy, and the third line follows because ζ is an N^T-dimensional state.

Now for all $i \in [K]$, if we let $\lambda_{i,1}, \ldots, \lambda_{i,N}$ be the eigenvalues of σ_i, half the $\lambda_{i,j}$'s are $\frac{1/2+3\varepsilon}{N/2}$ and the other half are $\frac{1/2-3\varepsilon}{N/2}$; this can be seen by applying a unitary transformation that diagonalizes σ_i, by rotating to a basis that contains $\frac{N}{2}$ basis vectors for S_i. Hence

$$\begin{aligned}
\mathrm{S}(\sigma_i) &= \sum_{x=1}^{N} \lambda_{i,x} \log_2 \frac{1}{\lambda_{i,x}} \\
&= \frac{N}{2}\left(\frac{1/2+3\varepsilon}{N/2}\right)\log_2\left(\frac{N/2}{1/2+3\varepsilon}\right) + \frac{N}{2}\left(\frac{1/2-3\varepsilon}{N/2}\right)\log_2\left(\frac{N/2}{1/2-3\varepsilon}\right) \\
&= \log_2 N - \left[1 - \left(\frac{1}{2}+3\varepsilon\right)\log_2\left(\frac{1}{1/2+3\varepsilon}\right) - \left(\frac{1}{2}-3\varepsilon\right)\log_2\left(\frac{1}{1/2-3\varepsilon}\right)\right] \\
&\geq \log_2 N - O\left(\varepsilon^2\right).
\end{aligned}$$

Combining,

$$\mathrm{I}(\zeta; i) = O\left(T\varepsilon^2\right).$$

We conclude that, in order to achieve quantum mutual information $\mathrm{I}(\zeta; i) = \Omega\left(N^2\right)$, so that ζ can determine i,

$$T = \Omega\left(\frac{N^2}{\varepsilon^2}\right) = \Omega\left(\frac{\min\left\{D^2, \log M\right\}}{\varepsilon^2}\right)$$

copies of σ_i are information-theoretically necessary. ■

Let's also observe a case in which Theorem 19 has a matching upper bound.

Proposition 20 *Given an unknown mixed state ρ, and known two-outcome measurements $E_1, \ldots, E_M$ and reals $c_1, \ldots, c_M \in [0,1]$, suppose we're promised that for each $i \in [M]$, either $\operatorname{Tr}(E_i \rho) \geq c_i$ or $\operatorname{Tr}(E_i \rho) \leq c_i - \varepsilon$. Then we can decide which is the case for each $i \in [M]$ using $k = O\left(\frac{\log M/\delta}{\varepsilon^2}\right)$ copies of ρ, with success probability at least $1-\delta$.*

Proof. For each $i \in [M]$, we simply perform a collective measurement on $\rho^{\otimes k}$, which applies E_i to each copy of ρ, and accepts if and only if the number of accepting invocations is at least $\left(c_i - \frac{\varepsilon}{2}\right)k$. By a Chernoff bound, this causes us to learn the truth for that i with probability at least

$$1 - \exp\left(-\varepsilon^2 \frac{\log M/\delta}{\varepsilon^2}\right) \geq 1 - \frac{\delta^2}{4M^2},$$

provided the constant in the big-O is sufficiently large. By Lemma 12 (the Quantum Union Bound), this means that by applying these collective measurements successively, we can learn the truth for *every* $i \in [M]$ with probability at least

$$1 - 2M\sqrt{\frac{\delta^2}{4M^2}} = 1 - \delta.$$

■

Of course, we also know from the recent work of O'Donnell and Wright [27] and Haah et al. [18] that $O\left(\frac{D^2}{\varepsilon^2}\right)$ copies of ρ suffice to learn ρ up to ε error in trace distance, so that number suffices for shadow tomography as well.

7 Open Problems

This paper initiated the study of shadow tomography of quantum states, and proved a surprising upper bound on the number of copies of a state that suffice for it. But this is just the beginning of what one can ask about the problem. Here we discuss four directions for future work.

(1) Tight Bounds. What is the true sample complexity of shadow tomography? We conjecture that Theorem 2 is far from tight. Our best current lower bound, Theorem 19, is $\Omega\left(\frac{\min\left\{D^2,\log M\right\}}{\varepsilon^2}\right)$. Could shadow tomography be possible with only $\left(\frac{\log M}{\varepsilon}\right)^{O(1)}$ copies of ρ, independent of D—as it is in the classical case (by Proposition 17), and the case of a promise gap (by Proposition 20)? Can we prove any lower bound of the form $\omega(\log M)$?

Also, what happens if we consider measurements with $K > 2$ outcomes? In that setting, it's easy to give *some* upper bound for the state complexity of shadow tomography, by reducing to the two-outcome case. Can we do better?

One can also study the state complexity of many other learning tasks. For example, what about what we called the "gentle search problem": finding an $i \in [M]$ for which $\operatorname{Tr}(E_i\rho)$ is large, promised that such an i exists? Can we improve our upper bound of $\widetilde{O}\left(\frac{\log^4 M}{\varepsilon^2}\right)$ copies of ρ for that task? Or what about approximating the vector of $\operatorname{Tr}(E_i\rho)$'s in other norms, besides the ∞-norm?

(2) Shadow Tomography with Restricted Kinds of Measurements. From an experimental standpoint, there are at least three drawbacks of our shadow tomography procedure. First, the procedure requires collective measurements on roughly $\frac{\log D}{\varepsilon^2}$ copies of ρ, rather than measurements on each copy separately (or at any rate, collective measurements on a smaller number of copies, like $\log\log D$). Second, the procedure requires so-called *non-demolition measurements*, which carefully maintain a state across a large number of sequential measurements (or alternatively, but just as inconveniently for experiments, an extremely long circuit to implement a single measurement). Third, the actual measurements performed are not just amplified E_i's, but the more complicated measurements required by Harrow et al. [19].

It would be interesting to address these drawbacks, either alone or in combination. To illustrate, if one could prove the soundness of Aaronson's original procedure for the Quantum OR Bound [3], that would remove the third drawback, though not the other two.

(3) Computational Efficiency. To what extent can our results be made *computationally* efficient, rather than merely efficient in the number of copies of ρ? In Section 1.3, we estimated the computational complexity of our shadow tomography procedure as $\widetilde{O}(ML) + D^{O(\log\log D)}$ (ignoring the dependence on ε and δ), where L is the length of a circuit to implement a single E_i. We also discussed the recent work of Brandão et al. [14], which builds on this work to do shadow tomography in $\widetilde{O}\left(\sqrt{M}L\right) + D^{O(1)}$ time, again using $\operatorname{poly}(\log M, \log D)$ copies of ρ—or in $\widetilde{O}\left(\sqrt{M} \operatorname{polylog} D\right)$ time under strong additional assumptions about the E_i's.

How far can these bounds be improved, with or without extra assumptions on ρ or the E_i's?

There are some obvious limits. If the measurements $E_1, \ldots, E_M$ are given by $D \times D$ Hermitian matrices (or by circuits of size D^2), and if the algorithm first needs to load descriptions of all the measurements into memory, then that already takes $\Omega\left(MD^2\right)$ time, or $\Omega\left(\sqrt{M}\right)$ time just to Grover search among the measurements. Applying a measurement with circuit complexity D^2 takes D^2 time. Outputting estimates for each $\operatorname{Tr}(E_i \rho)$ takes $\Omega(M)$ time—although Brandão et al. [14] evade that bound by letting their output take the form of a quantum circuit to prepare a state σ such that $|\operatorname{Tr}(E_i \sigma) - \operatorname{Tr}(E_i \rho)| \le \varepsilon$ for all $i \in [M]$, which seems fair.

If we hope to do even better than that, we could demand that the measurements E_i be implementable by a *uniform* quantum algorithm, which takes i as input and runs in time $\operatorname{polylog} D$. Let's call a shadow tomography procedure *hyperefficient* if, given as input such a uniform quantum algorithm A, as well as $k = \operatorname{poly}\left(\log M, \log D, \frac{1}{\varepsilon}\right)$ copies of ρ, the procedure uses $\operatorname{poly}\left(\log M, \log D, \frac{1}{\varepsilon}\right)$ time to output a quantum circuit C such that $|C(i) - \operatorname{Tr}(E_i \rho)| \le \varepsilon$ for all $i \in [M]$. Note that, in the special case that ρ and the E_i's are classical, hyperefficient shadow tomography is actually possible, since we can simply output a circuit C that hardwires k classical samples $x_1, \ldots, x_k \sim \mathcal{D}$, and then on input i, returns the empirical mean of $E_i(x_1), \ldots, E_i(x_k)$.

In the general case, by contrast, we observe the following:

Proposition 21 *Suppose there exists a hyperefficient shadow tomography procedure. Then quantum advice can always be simulated by classical advice—i.e.,* $\mathsf{BQP/qpoly} = \mathsf{BQP/poly}$.

Proof. This already follows from the assumption that a quantum circuit C of size $\operatorname{poly}(\log M, \log D)$, satisfying (say) $|C(i) - \operatorname{Tr}(E_i \rho)| \le \frac{1}{10}$ for every $i \in [M]$, *exists.* For a description of that C can be provided as the $\mathsf{BQP/poly}$ advice when simulating $\mathsf{BQP/qpoly}$. ■

Note that Aaronson and Kuperberg [11] gave a quantum oracle relative to which $\mathsf{BQP/qpoly} \neq \mathsf{BQP/poly}$. By combining that with Proposition 21, we immediately obtain a quantum oracle relative to which hyperefficient shadow tomography is impossible.

One further observation:

Proposition 22 *Suppose there exists a hyperefficient shadow tomography procedure. Then quantum copy-protected software (see [5] or Section 2) is impossible.*

Proof. Given $n^{O(1)}$ copies of a piece ρ of quantum software, by assumption we could efficiently produce an $n^{O(1)}$-bit classical string s, which could be freely copied and would let a user efficiently compute ρ's output behavior (i.e., accepting or rejecting) on any input $x \in \{0,1\}^n$ of the user's choice. ■

It would be interesting to know what further improvements are possible to the computational complexity of shadow tomography, consistent with the obstacles mentioned above. Also, even if shadow tomography inherently requires exponential computation time in the worst case, one naturally seeks do better in special cases. For example, what if the E_i's are stabilizer measurements? Or if ρ is a low-dimensional matrix product state?

(4) Applications. A final, open-ended problem is to *find more applications* of shadow tomography. In Section 2, we gave implications for quantum money, quantum software, quantum advice, and quantum communication protocols. But something this basic being possible seems like it ought to have further applications in quantum information theory, and conceivably even experiment. In the search for such applications, we're looking for any situation where (i) one has an unknown entangled state ρ on many particles; (ii) one is limited mainly in how many copies of ρ one can produce; (iii) one wants to know, at least implicitly, the approximate expectation values of ρ on a huge number of observables; and (iv) one doesn't need to know more than that.

8 Acknowledgments

I thank Fernando Brandão, Patrick Hayden, Cedric Lin, Guy Rothblum, and Ronald de Wolf for helpful discussions and comments, Steve Flammia for suggesting the name "shadow tomography," and the anonymous reviewers for their suggestions.

References

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

[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. Conference on Computational Complexity*, pages 261–273, 2006. quant-ph/0510230.

[4] S. Aaronson. The learnability of quantum states. *Proc. Roy. Soc. London*, A463(2088):3089–3114, 2007. quant-ph/0608142.

[5] S. Aaronson. Quantum copy-protection and quantum money. In *Proc. Conference on Computational Complexity*, pages 229–242, 2009. arXiv:1110.5353.

[6] S. Aaronson. The complexity of quantum states and transformations: From quantum money to black holes, February 2016. Lecture Notes for the 28th McGill Invitational Workshop on Computational Complexity, Holetown, Barbados. With guest lectures by A. Bouland and L. Schaeffer. www.scottaaronson.com/barbados-2016.pdf.

[7] S. Aaronson, X. Chen, E. Hazan, S. Kale, and A. Nayak. Online learning of quantum states. In *Proc. of Neural Information Processing Systems (NIPS)*, 2018. arXiv:1802.09025.

[8] S. Aaronson, X. Chen, E. Hazan, and A. Nayak. Online learning of quantum states. arXiv:1802.09025, 2018.

[9] S. Aaronson and P. Christiano. Quantum money from hidden subspaces. *Theory of Computing*, 9:349–401, 2013. Earlier version in STOC'2012. arXiv:1203.4740.

[10] S. Aaronson and A. Drucker. A full characterization of quantum advice. *SIAM J. Comput.*, 43(3):1131–1183, 2014. Earlier version in STOC'2010. arXiv:1004.0377.

[11] S. Aaronson and G. Kuperberg. Quantum versus classical proofs and advice. *Theory of Computing*, 3(7):129–157, 2007. Earlier version in CCC'2007. arXiv:quant-ph/0604056.

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

[13] C. H. Bennett, G. Brassard, S. Breidbart, and S. Wiesner. Quantum cryptography, or unforgeable subway tokens. In *Proceedings of CRYPTO*, pages 267–275. Plenum Press, 1982.

[14] F. Brandão, A. Kalev, T. Li, C. Lin, K. Svore, and X. Wu. Exponential quantum speed-ups for semidefinite programming with applications to quantum learning. arXiv:1710.02581, 2017.

[15] M. Ettinger, P. Høyer, and E. Knill. The quantum query complexity of the hidden subgroup problem is polynomial. *Inform. Proc. Lett.*, 91(1):43–48, 2004. quant-ph/0401083.

[16] E. Farhi, D. Gosset, A. Hassidim, A. Lutomirski, D. Nagaj, and P. Shor. Quantum state restoration and single-copy tomography. *Phys. Rev. Lett.*, 105(190503), 2010. arXiv:0912.3823.

[17] D. Gross, Y. Liu, S. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(150401), 2010. arXiv:0909.3304.

[18] J. Haah, A. Harrow, Z. Ji, X. Wu, and N. Yu. Sample-optimal tomography of quantum states. *IEEE Trans. Information Theory*, 63(9):5628–5641, 2017. Earlier version in STOC'2016. arXiv:1508.01797.

[19] A. Harrow, C. Lin, and A. Montanaro. Sequential measurements, disturbance and property testing. In *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 1598–1611, 2017. arXiv:1607.03236.

[20] P. Hayden, D. W. Leung, and A. Winter. Aspects of generic entanglement. *Communications in Mathematical Physics*, 265(1):95–117, 2006. quant-ph/0407049.

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

[22] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In *Proc. of Computational Learning Theory (COLT)*, pages 1066–1100, 2015.

[23] C. Marriott and J. Watrous. Quantum Arthur-Merlin games. *Computational Complexity*, 14(2):122–152, 2005. Earlier version in CCC'2004. arXiv:cs/0506068.

[24] A. Molina, T. Vidick, and J. Watrous. Optimal counterfeiting attacks and generalizations for Wiesner's quantum money. In *Theory of Quantum Computation, Communication, and Cryptography*, pages 45–64, 2012. arXiv:1202.4010.

[25] A. Montanaro. Learning stabilizer states by Bell sampling. arXiv:1707.04012, 2017.

[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] R. O'Donnell and J. Wright. Efficient quantum tomography. In *Proc. ACM STOC*, pages 899–912, 2016. arXiv:1508.01907.

[28] C. Song, K. Xu, W. Liu, C. Yang, S.-B. Zheng, H. Deng, Q. Xie, K. Huang, Q. Guo, L. Zhang, P. Zhang, D. Xu, D. Zheng, X. Zhu, H. Wang, Y.-A. Chen, C.-Y. Lu, S. Han, and J.-W. Pan. 10-qubit entanglement and parallel logic operations with a superconducting circuit. *Phys. Rev. Lett.*, 119(180511), 2017. arXiv:1703.10302.

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

[30] S. Wiesner. Conjugate coding. *SIGACT News*, 15(1):78–88, 1983. Original manuscript written circa 1970.

[31] M. Wilde. Sequential decoding of a general classical-quantum channel. *Proc. Roy. Soc. London*, A469(2157):20130259, 2013. arXiv:1303.0808.

[32] A. Winter. Coding theorem and strong converse for quantum channels. *IEEE Trans. Information Theory*, 45(7):2481–2485, 1999. arXiv:quant-ph/0012127.