Gentle Measurement of Quantum States and Differential Privacy

Scott Aaronson* Guy N. Rothblum†

Abstract

In *differential privacy (DP)*, we want to query a database about n users, in a way that "leaks at most ε about any individual user," even conditioned on any outcome of the query. Meanwhile, in *gentle measurement*, we want to measure n quantum states, in a way that "damages the states by at most α," even conditioned on any outcome of the measurement. In both cases, we can achieve the goal by techniques like deliberately adding noise to the outcome before returning it. This paper proves a new and general connection between the two subjects. Specifically, we show that on products of n quantum states, any measurement that is α-gentle for small α is also $O(\alpha)$-DP, and any product measurement that is ε-DP is also $O(\varepsilon\sqrt{n})$-gentle.

Illustrating the power of this connection, we apply it to the recently studied problem of *shadow tomography*. Given an unknown d-dimensional quantum state ρ, as well as known two-outcome measurements $E_1, \ldots, E_m$, shadow tomography asks us to estimate $\Pr[E_i \text{ accepts } \rho]$, for *every* $i \in [m]$, by measuring few copies of ρ. Using our connection theorem, together with a quantum analog of the so-called *private multiplicative weights* algorithm of Hardt and Rothblum, we give a protocol to solve this problem using $O\left((\log m)^2 (\log d)^2\right)$ copies of ρ, compared to Aaronson's previous bound of $\widetilde{O}\left((\log m)^4 (\log d)\right)$. Our protocol has the advantages of being *online* (that is, the E_i's are processed one at a time), gentle, and conceptually simple.

Other applications of our connection include new *lower* bounds for shadow tomography from lower bounds on DP, and a result on the safe use of estimation algorithms as subroutines inside larger quantum algorithms.

Contents

*University of Texas at Austin. Email: aaronson@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.

†Weizmann Institute of Science. Email: rothblum@alum.mit.edu. Supported by ISF grant no. 5219/17.

1 Introduction

This paper is about a new mathematical connection between two concepts—*gentle measurement* in quantum mechanics, and *differential privacy* in classical computer science—and the applications of this connection to the design of new quantum measurement procedures and algorithms. Since the paper is meant to be accessible to researchers in both fields (and beyond), we begin by saying a few words about each of the concepts separately.

1.1 Gentle Measurement

In quantum mechanics, *measurement* is, famously, an inherently destructive process. For example, if we measure a qubit $\alpha|0\rangle + \beta|1\rangle$ in the $\{|0\rangle, |1\rangle\}$ basis, we "force the qubit to decide" whether to be $|0\rangle$ (with probability $|\alpha|^2$) or $|1\rangle$ (with probability $|\beta|^2$). The qubit's state then "collapses" to whichever choice it made. There's no way to measure again, unless of course we happen to have (or know how to prepare) a second qubit in the same state.[1]

Even in quantum mechanics, though, not all measurements on all states are destructive. For example, if a qubit happens to be in the $|0\rangle$ state already, then measuring in the $\{|0\rangle, |1\rangle\}$ basis causes no damage at all. And if the qubit is the state $|\psi\rangle = \sqrt{1-\varepsilon^2}|0\rangle + \varepsilon|1\rangle$ for small ε, then measuring in the $\{|0\rangle, |1\rangle\}$ basis causes only minimal damage, since the result is almost always that the qubit "snaps" to $|0\rangle \approx |\psi\rangle$. More generally, the principle is this:

> A measurement M applied to a state $|\psi\rangle$ necessarily severely damages $|\psi\rangle$ if, and only if, the outcome of M is highly unpredictable even to someone who already knows $|\psi\rangle$.

This principle, which can be quantified in various ways, is called *information/disturbance tradeoff*: if M creates lots of new (random) information, then it must also cause lots of disturbance to $|\psi\rangle$, and vice versa.

A corollary is that, if someone who knew $|\psi\rangle$ could usually predict the measurement outcome in advance, then applying M need not damage $|\psi\rangle$ by much. Note that this corollary does not describe only trivial or uninteresting measurements, since in general the measurer does *not* know $|\psi\rangle$ in advance—that's why she's measuring it![2]

Indeed, so-called *gentle measurements*, which can be limited in how much damage they cause, have found numerous applications in experimental physics, the foundations of quantum mechanics, and quantum computing theory.[3] Experimentalists, for example, know how to perform a measurement on a large number of identically prepared particles, in a way that reveals the particles'

[1] This destructiveness is not unique to quantum mechanics: it has a close analogue in classical Bayesian conditioning, where a probability distribution can "collapse" to a single point when we make an observation. But in classical probability, the "collapse" is purely internal and mental, in the sense that we could undo it by simply forgetting the observation. In quantum mechanics, by contrast, collapse causes an *objective* change to the measured system, one that could also be detected by someone else who later measured the system.

[2] And also, even if she *did* know a description of $|\psi\rangle$, she might still find predicting the outcome of a measurement on $|\psi\rangle$ to be computationally intractable.

[3] Physicists more often refer to "weak measurement," a related but not identical concept, which typically means that the measurement returns very little information about the state (in this paper, we'll call such measurements "ε-trivial"). All weak measurements can be implemented gently, and we'll show in Lemma 23 that the only measurements that are gentle on *all* states are weak. But measurements that are gentle on large sets of interesting states (such as product states) can be far from weak, a point that will be crucial for us.

quantum state to high accuracy while causing very little damage.[4] More theoretically, gentle measurement has also played a central role in proposals for *publicly-verifiable quantum money* that can be verified many times, *quantum software* that can be evaluated on many inputs, and so forth (see [4, 6]). Gentle measurement is also needed in work on the nonabelian hidden subgroup problem [24], and on quantum advice complexity classes like BQP/qpoly (see [2, 3]).

Let's now define a bit more formally what we'll mean, in this paper, by a quantum measurement being "gentle."

Definition 1 (Gentle Measurements) *Given a set S of quantum mixed states in some Hilbert space, an implementation of a measurement M,[5] and a parameter $\alpha \in [0,1]$, we define M to be* ***α-gentle on*** *S if for all states $\rho \in S$, and all possible outcomes y of applying M to ρ, we have*

$$\left\|\rho_{M \to y} - \rho\right\|_{\text{tr}} \leq \alpha. \tag{1}$$

Here $\|\cdot\|_{\text{tr}}$ represents trace distance, the standard distance metric on quantum states, while $\rho_{M \to y}$ represents the new, "collapsed" state assuming that the measurement outcome was y. (For a review of these and other quantum information concepts, see Sections 2.2 and 2.5.)

More generally, we say M is (α, δ)-gentle on S if for all states $\rho \in S$, inequality (1) holds with probability at least $1-\delta$ over the possible outcomes y of applying M to ρ. We recover α-gentleness by setting $\delta = 0$.

The most common choices for S will be the set of product states $\rho = \rho_1 \otimes \cdots \otimes \rho_n$, and the set of all states.

If a measurement M is specified by its output probabilities only (technically, as a "POVM"), then we say that M is α-gentle if and only if there exists an α-gentle implementation of it.

As an example, suppose we have n qubits in a pure product state:

$$|\psi\rangle = (\alpha_1 |0\rangle + \beta_1 |1\rangle) \otimes \cdots \otimes (\alpha_n |0\rangle + \beta_n |1\rangle).$$

Then consider the measurement M that simply returns the total Hamming weight. This measurement is *not* α-gentle for any nontrivial α. So for example, if we apply M to the equal superposition $\left(\frac{|0\rangle+|1\rangle}{\sqrt{2}}\right)^{\otimes n}$, we'll collapse the superposition over possible Hamming weights—from a Gaussian wavepacket (as the physicists might call it) of width $\Theta(\sqrt{n})$ centered at $n/2$, all the way down to a single random Hamming weight.

By contrast, now consider a measurement L_σ that returns the Hamming weight, *plus* a random noise term η of average magnitude $\sigma \gg \sqrt{n}$. As an example, we could take this noise to follow a Laplace distribution:

$$\Pr[\eta = k] = \frac{1}{Z} e^{-|k|/\sigma}, \tag{2}$$

where

$$Z = \frac{2}{1 - e^{-1/\sigma}} - 1 \approx 2\sigma - 1$$

[4] With a single particle, this is of course impossible.

[5] In this paper, by an "implementation" of a measurement M, we mean a specification from which, given a state ρ, one can calculate not only the probabilities of the various outcomes y, but *also* the post-measurement states $\rho_{M \to y}$.

for large σ. We can implement the measurement L_σ as follows. Given $|\psi\rangle$, which we now think of as a superposition $\sum_{X\in\{0,1\}^n} \alpha_X |X\rangle$ over n-bit strings, first prepare alongside $|\psi\rangle$ the state

$$|\eta\rangle := \sum_{k=-\infty}^{\infty} \sqrt{\Pr[\eta = k]}\,|k\rangle\,.$$

(In practice, we would of course impose a cutoff on $|k|$.) Next, perform the unitary transformation

$$\sum_{X\in\{0,1\}^n}\sum_{k=-\infty}^{\infty} \alpha_X \sqrt{\Pr[\eta = k]}\,|X\rangle\,|k\rangle \rightarrow \sum_{X\in\{0,1\}^n}\sum_{k=-\infty}^{\infty} \alpha_X \sqrt{\Pr[\eta = k]}\,|X\rangle\,||X| + k\rangle\,.$$

Finally, measure the $||X| + k\rangle$ register in the standard basis and output the result.

It turns out that this noisy measurement L_σ is $O(\sqrt{n}/\sigma)$-gentle.[6] Intuitively, this is because the various Hamming weights that are well-represented in the "Gaussian wavepacket" $|\psi\rangle$—e.g., in the example $|\psi\rangle = \left(\frac{|0\rangle+|1\rangle}{\sqrt{2}}\right)^{\otimes n}$, those Hamming weights w such that $\left|w - \frac{n}{2}\right| = O(\sqrt{n})$—lead to probability distributions over measurement outcomes that mostly overlap. In other words, when we observe an outcome of the form $|X| + k$, the intrinsic variation in $|X|$ within the superposition is dominated by the variation in k.

1.2 Differential Privacy

Differential privacy (DP) is a young subfield of computer science—younger than quantum computing, actually—that's seen tremendous growth since its beginnings around 2006 [20, 21, 44]. Though as we'll see, DP's concepts turn out to have much broader applicability, the original motivating problem is as follows. Suppose that a hospital (or bank, or social media site) has a database of sensitive personal records. The hospital wants to let medical researchers query its database in such a way that

(1) the researchers can learn as much accurate statistical information as possible about the patient population (e.g., how many of them have colon cancer), but

(2) each patient has a mathematical guarantee that, by opting to participate in the database, she's exposing to the researchers "only a negligible amount" of data about herself that would otherwise be private.

The question is, how should we design the queries to balance these two apparently conflicting demands?

More formally, call two databases X, X' *neighbors* if they differ only in a single entry x_i. Then here is the key definition:

Definition 2 (Differential Privacy [20]) *Given a randomized algorithm A that queries a database X, as well as a parameter $\varepsilon \geq 0$, we define A to be ε-**DP** if for all databases X, X' that are neighbors, and all possible outputs y of A, we have*

$$\Pr[A(X) = y] \leq e^{\varepsilon} \Pr\left[A\left(X'\right) = y\right].$$

Here the probabilities are over the internal randomness used by A.

[6] While there are other ways to prove that L_σ is $O(\sqrt{n}/\sigma)$-gentle, the nicest proof we know will deduce it as an immediate corollary of this paper's main results.

In place of e^ε, one could also use the more intuitive $1+\varepsilon$. However, the choice of e^ε has the advantages that it composes nicely and is symmetric under inversions.

As an example—which should look familiar!—suppose the databases X are n-bit strings, and consider the algorithm that simply returns the Hamming weight $|X|$. This algorithm is *not* ε-DP for any ε, since flipping just a single bit of X can change the probability of an output (namely, the new Hamming weight) from 0 to 1. By contrast, now consider the algorithm L_σ that returns the Hamming weight $|X|$, *plus* a Laplace noise term η that's distributed according to equation (2). For any two neighboring databases X, X', and any possible output y, we have

$$\frac{\Pr\left[L_\sigma\left(X\right)=y\right]}{\Pr\left[L_\sigma\left(X'\right)=y\right]}=\frac{e^{-|y-|X||/\sigma}}{e^{-|y-|X'||/\sigma}}\leq e^{1/\sigma}. \tag{3}$$

So we see that L_σ is $\frac{1}{\sigma}$-DP. Yet, as long as σ is not too enormous, the output $|X|+\eta$ still provides a useful estimate of $|X|$.

Requiring multiplicative closeness in the probabilities of every output y might seem overly strong. But if we relaxed the definition to an additive one, we'd need to admit the algorithm that simply chooses a user $i\in[n]$ uniformly at random and publishes all of her data. This algorithm is manifestly not "private," and yet it satisfies a strong additive guarantee: if user i changes her data, that will affect the probability distribution over outputs by at most $\frac{1}{n}$ in variation distance. On the other hand, one can check that this algorithm is not ε-DP for any finite ε.

DP has been applied in deployed systems, for example at Apple and Google; see for example [42] for discussion. The concept has also found application to other problems, not obviously related to privacy—for example, adaptive data analysis (for more see Section 1.7). But what does DP have to do with quantum information in general, or gentle measurement in particular?

1.3 The Connection

Given two quantum mixed states ρ,σ on n registers each, call them *neighbors* if it's possible to reach either σ from ρ, or ρ from σ, by performing a general quantum operation (a so-called superoperator) on a single register only. In the special case where $\rho=\rho_1\otimes\cdots\otimes\rho_n$ and $\sigma=\sigma_1\otimes\cdots\otimes\sigma_n$ are both product states, this reduces to saying: ρ and σ are neighbors if and only if $\rho_i\neq\sigma_i$ for at most one i.

Using this notion, we can easily generalize the definition of differential privacy from Section 1.2 to the quantum setting:

Definition 3 (Quantum Differential Privacy) *Given a set S of quantum mixed states each on n registers, a measurement M, and a parameter $\varepsilon\geq 0$, we define M to be ε-**DP on** S if for all states $\rho,\sigma\in S$ that are neighbors, and all possible outputs y of M, we have*

$$\Pr\left[M\left(\rho\right)=y\right]\leq e^\varepsilon\Pr\left[M\left(\sigma\right)=y\right]. \tag{4}$$

Here the probabilities are over the intrinsic randomness of the measurement outcome.

More generally, we say M is (ε,δ)-DP on S if for all neighboring states $\rho,\sigma\in S$, inequality (4) holds with probability at least $1-\delta$ over the possible outcomes y of applying M to ρ. We recover ε-DP by setting $\delta=0$.[7]

[7] This is a slightly nonstandard definition of (ε,δ)-DP, but can be related to the standard definition by a nontrivial result. See e.g. Vadhan [44, Lemma 1.5].

The most common choices for S will be the set of product states $\rho = \rho_1 \otimes \cdots \otimes \rho_n$, and the set of all states.

Note that unlike with gentleness, the property of being ε-DP depends only on the output probabilities, and not at all on the post-measurement states (i.e., on the "implementation" of the measurement).

Perhaps the first question we should ask is: *are* there any nontrivial quantum measurements that satisfy the above definition? Indeed there are.

Recall the DP algorithm L_σ from Section 1.2, which returns the Hamming weight $|X|$ of an n-bit input database X, plus Laplace noise η of average magnitude σ. We can promote L_σ to a quantum measurement on n-qubit states, by implementing it using the procedure described in Section 1.1. We then have the following:

Proposition 4 *L_σ is $\frac{1}{\sigma}$-DP on the set of all n-qubit states.*

Proof. Since L_σ only involves measuring the Hamming weight in the computational basis, for any n-qubit state ρ we can write

$$\Pr\left[L_\sigma\left(\rho\right)=y\right]=\sum_{X\in\{0,1\}^n}\rho_{X,X}\Pr\left[L_\sigma\left(X\right)=y\right].$$

Also, if we act on a single register of ρ, and then measure in the computational basis (which by the above, we can do without loss of generality), we map each database X to a distribution over neighbors X' of X. The proposition now follows from convexity together with equation (3). ■

Stepping back, we've seen that simply measuring the Hamming weight of an n-qubit state is neither gentle nor private. And yet the same fix—namely, adding random noise to the measurement outcome before returning it—makes the measurement both gentle *and* private. Is this convergence, between gentle quantum measurement and differential privacy, just a coincidence?

Our main result asserts that it's not a coincidence: there's a strong two-way connection between the two notions.

Theorem 5 (Main Result) *For all quantum measurements M on n registers:*

(1) If M is α-gentle on product states for $\alpha \leq \frac{1}{4.01}$, then M is $O\left(\alpha\right)$-DP on product states.[8]

(2) If M is ε-DP on product states, and is a product measurement,[9] *then M is $O\left(\varepsilon\sqrt{n}\right)$-gentle on product states.*[10]

Again, here a "measurement" M corresponds to a specification of output probabilities; for M to be α-gentle means that there *exists* an α-gentle implementation of M.

Intuitively, it's far from obvious that gentleness and differential privacy should be connected in this way. After all, the definition of α-gentleness makes no reference to the notion of "neighboring" states. Conversely, the definition of ε-DP is exclusively concerned with output probabilities, and

[8] Indeed, it suffices for α to be bounded below $\frac{1}{4}$ by any fixed constant (which then affects the multiplier in the $O\left(\alpha\right)$). Similar remarks apply wherever constants like $\frac{1}{4.01}$ appear in this paper.

[9] That is, if M can be implemented by first applying a classical algorithm to the outcomes of separate POVM measurements on the n registers, and then uncomputing the outcomes of those n measurements.

[10] On non-product states, M will still produce the correct output probabilities, but it need not be gentle.

makes no reference to post-measurement states. Our goal is to explain why gentleness and DP *are* connected in this way, and to explore the consequences of the connection.

We'll see some applications of Theorem 5 shortly, in Sections 1.4 and 1.5. Before we do so, however, let's make a few comments about the theorem statement.

At first glance, part (2) of the theorem seems weaker than part (1)—especially because of the $\sqrt{n}$ blowup in parameters—but it's the part that carries many of the interesting implications. In Section 5, we'll show that the $\sqrt{n}$ blowup is unavoidable. Indeed the measurement L_σ, with $\sigma = \Theta(\sqrt{n})$, already demonstrates this.

By contrast, the condition that M is a product measurement is *not* clearly necessary; one of the central open problems we leave is whether that condition can be removed. In Appendix 12, we'll give examples of quantum DP measurements that can't be approximated by any product or (we conjecture) even LOCC measurements. However, all the examples we currently know of such measurements are extremely artificial.

The restriction to product states might seem strange, but it's provably unavoidable if we want Theorem 5 to say anything about nontrivial measurements. As we'll show in Section 3, there *is* a counterpart of Theorem 5 for states that could have arbitrary correlation or entanglement among the registers. It turns out, however, that if a measurement M is α-gentle on *all* states for $\alpha \ll \frac{1}{4}$, then M is close to trivial (i.e., it barely depends on the input state at all). And conversely, if M is ε-DP, then the best we can deduce is that M is $O(\varepsilon n)$-gentle on all states, rather than $O(\varepsilon\sqrt{n})$-gentle. While that might sound like a merely quantitative gap, the trouble is, again, that the only measurements that are ε-DP for $\varepsilon \ll \frac{1}{n}$ are close to trivial. By contrast, plenty of interesting measurements are ε-DP for $\varepsilon \ll \frac{1}{\sqrt{n}}$.

One might wonder whether our reductions between privacy and gentleness preserve computational efficiency. In one direction—turning gentleness into privacy—the answer is clearly yes, since an α-gentle measurement *is* $O(\alpha)$-DP; nothing further needs to be done. However, in the other direction—turning privacy into gentleness—we can implement a gentle measurement M efficiently only if we have an efficient algorithm to "QSample" M's output distribution on a given input. QSampling is a term coined in 2003 by Aharonov and Ta-Shma [9], which just means that we can efficiently prepare a superposition over outputs of the form

$$\sum_y \sqrt{\Pr[y]}\,|y\rangle,$$

which is not entangled with any "garbage" dependent on y. In practice, most DP algorithms that we know about *do* give rise to efficient QSampling procedures, but this property doesn't follow automatically from a DP algorithm's being efficient. In Section 7, we'll explore the issue of computational efficiency further, and give nontrivial conditions under which efficiency is preserved.

1.4 Applications

Can we *exploit* the connection between gentle measurement and differential privacy to port results from one field to the other, as was done with the connections between communication complexity and circuit lower bounds, cryptography and learning, etc.? The second main contribution of this paper is to use Theorem 5, together with previous work in DP, to derive new results in quantum measurement theory and quantum algorithms.[11]

[11] Some of these applications could also have been obtained by "brute force" (e.g., directly designing and analyzing the desired gentle measurements), but the connection to DP will both guide us to the correct statements, and enable

As a tiny warmup application, notice that L_σ, the Laplace noise measurement from Section 1.2, is a "product measurement," in the sense that it can be implemented via an algorithm that measures each register separately. And thus, by combining part (2) of Theorem 5 with Proposition 4, we immediately obtain the following:

Corollary 6 (Gentleness of Laplace Noise Measurement) *L_σ is $O\left(\sqrt{n}/\sigma\right)$ -gentle on product states.*

As far as we know, proving Corollary 6 directly would require a laborious calculation.

Here is another application. In the early days of quantum computing, Bennett et al. [11] observed that a quantum algorithm can safely invoke other quantum algorithms for decision problems as subroutines inside of a superposition—or in terms of complexity classes, that $\mathsf{BQP}^{\mathsf{BQP}} = \mathsf{BQP}$. The proof uses amplification, to push down the subroutine's error probability, combined with *uncomputing*, to eliminate any "garbage" that the subroutine leaves entangled with its input. However, this straightforward uncomputing strategy no longer works for subroutines whose purpose is to *estimate an expectation value to within $\pm\varepsilon$* (say, the acceptance probability of a quantum circuit).

In Section 7, we'll point out one simple solution to this problem: namely, *run the subroutine $n^{O(1)}$ times in parallel, then estimate the desired expectation values by simulating gentle measurements on the resulting states.* If we implement this idea using the Laplace noise measurement L_σ, then Corollary 6 yields the following:

Theorem 7 *Without loss of generality, a BQP algorithm can at any point estimate $\Pr[C \text{ accepts}]$ to within $\pm\frac{1}{n^{O(1)}}$, on any superposition containing descriptions of quantum circuits C, while maintaining the superposition's coherence.*

While it's possible to prove Theorem 7 "bare-handedly," without knowing about the connection between gentleness and DP, the point is that the floodgates are now open. Given a quantum algorithm P, which is run as a subroutine inside a larger quantum algorithm Q, there are many things that Q might want to know about P's output behavior, beyond just additive estimates for specific probabilities. Whatever the details, Theorem 5 reduces the task to designing a suitable efficient DP algorithm, or finding such an algorithm in the literature. Gentleness then follows automatically.

1.5 Shadow Tomography

In Section 6, we present our "flagship" application for the connection between gentleness and DP: a new quantum measurement procedure, called *Quantum Private Multiplicative Weights (QPMW)*, which achieves parameters and properties that weren't previously known.

QPMW addresses a task that Aaronson [5], in 2016, called *shadow tomography*. Here we're given n copies of an unknown d-dimensional mixed state ρ. We're also given known two-outcome measurements $E_1, \ldots, E_m$. Our goal is to learn $\Pr[E_i(\rho) \text{ accepts}]$ to within an additive error of $\pm\varepsilon$, for *every* $i \in [m]$, with high success probability (say, at least 2/3), by carefully measuring the

the simplest proofs of them that we know. Meanwhile, our applications to so-called *shadow tomography* of quantum states, described in Section 1.5, will make essential use of sophisticated algorithms and lower bounds from the DP literature.

ρ's. Setting aside computational difficulty, how many copies of ρ are information-theoretically necessary for this?

At one extreme of parameters, and suppressing the dependence on ε, it's clear that $n = \widetilde{O}(m)$ copies of ρ suffice, since we could just apply each E_i to different copies. At a different extreme, it's also clear that $n = \widetilde{O}\left(d^2\right)$ copies suffice—or not "clear," but it follows from celebrated recent work by O'Donnell and Wright [36] and (independently) Haah et al. [26], who showed that $\widetilde{O}\left(d^2\right)$ copies of ρ are necessary and sufficient for full *quantum state tomography*: that is, reconstructing the entire state ρ to suitable precision.

But what if we only want to learn the "shadow" that ρ casts on the measurements $E_1, \ldots, E_m$? Aaronson [5] raised the question of whether shadow tomography might be possible using a number of copies n that scales only polylogarithmically in both m and d—so in particular, that's polynomial even if m and d are both exponential. While this seemed overly ambitious, Aaronson was unable to rule it out; and indeed, last year he showed:

Theorem 8 (Aaronson [6]) *There exists an explicit procedure to perform shadow tomography using*

$$n = \widetilde{O}\left(\frac{(\log m)^4 (\log d)}{\varepsilon^4}\right)$$

copies of ρ. Here the $\widetilde{O}$ hides factors of $\log\log m$, $\log\log d$, and $\log\frac{1}{\varepsilon}$.

Shortly afterward, Brandão et al. [14] gave a different shadow tomography procedure, based on semidefinite programming, which achieved the same sample complexity as Aaronson's but was more efficient computationally.

However, these developments left several questions open:

(1) What is the true sample complexity of shadow tomography? The best lower bound in [6] is that $\Omega\left(\frac{\min\left\{d^2, \log m\right\}}{\varepsilon^2}\right)$ copies are needed.

(2) The procedures of [6, 14] destroy the copies of ρ in the process of measuring them. Is there a shadow tomography procedure that's also *gentle*?

(3) The procedures of [6, 14] require the full list $E_1, \ldots, E_m$ to be known in advance. Is there a shadow tomography procedure that's *online*—i.e., that receives the measurements one by one, and estimates each $\Pr[E_i \text{ accepts } \rho]$ immediately after receiving E_i?

In Section 6, by exploiting our connection between gentleness and DP, and by quantizing a known classical DP algorithm called Private Multiplicative Weights [27], we prove a new shadow tomography theorem that addresses all of the above questions.

Theorem 9 (Quantum PMW) *There exists an explicit procedure, Quantum Private Multiplicative Weights (QPMW), that performs shadow tomography with success probability $1 - \beta$ using*

$$n = O\left(\frac{\left(\log^2 m + \log\frac{1}{\delta}\right) \cdot \log^2 d \cdot \log\frac{1}{\beta}}{\varepsilon^8}\right)$$

copies of ρ, and which is also online and (ε, δ)-gentle.

Most notably, QPMW is both *online* and *gentle*; the previous procedures [6, 14] were neither. Because of its simplicity and its online nature, QPMW seems better suited than its predecessors to potential experimental realization.

Meanwhile, compared to Theorem 8, Theorem 9 improves the dependence on m from $(\log m)^4$ to $(\log m)^2$. The dependence on d and $1/\varepsilon$ is worse, but we conjecture that this is an artifact of our analysis, and that porting so-called "advanced composition" [22] to the quantum setting would ameliorate the situation. The running time of QPMW is roughly $O(mL) + d^{O(1)}$, where L is the time needed to implement a single E_i; this improves on the $O(mL) + d^{O(\log\log d)}$ running time of Aaronson's procedure, and matches an improvement from $d^{O(\log\log d)}$ to $d^{O(1)}$ in Brandão et al. [14].

It's hard to give a simple intuition for the improvement in m-dependence from $O\left(\log^4 m\right)$ to $O\left(\log^2 m\right)$. Loosely, though, gentleness (derived from DP) lets QPMW be online, and being online lets QPMW avoid the "gentle search procedure," a key subroutine in Aaronson's earlier procedure [6] that was responsible for the $\log^4 m$ factor. In any case, we wish to stress that quantitative improvements in sample complexity are not the main point here. The point, rather, is that the connection between DP and gentleness leads to an entirely new approach to shadow tomography.

The DP/gentleness connection turns out to be useful not just for *upper* bounds on the sample complexity of shadow tomography, but also lower bounds. In Section 6.3, we'll combine a recent lower bound on DP algorithms [16] with *part (1)* of Theorem 5 (i.e., the gentleness implies DP direction), to deduce a new lower bound on the sample complexity of gentle shadow tomography, where here "gentle" means "gentle on all product state inputs." We'll also use recent work from adaptive data analysis [35] to observe a lower bound on the sample complexity of *online* shadow tomography—showing that, for the latter task, QPMW's sample complexity is optimal up to polynomial factors.

Finally, in Section 7.2, we prove lower bounds on the *computational* complexity of gentle and online shadow tomography, by deducing them as corollaries of recent lower cryptographic bounds for differential privacy and adaptive data analysis [43, 28, 40]. Assuming the existence of a one-way function that takes $2^{\Omega(n)}$ time to invert, these lower bounds say that any algorithm for online or gentle shadow tomography needs $d^{\Omega(1)}$ time, so in that respect the QPMW procedure is optimal for those tasks.

We stress that all our lower bounds for shadow tomography—both information-theoretic and computational—are obtained by using this paper's machinery to port known classical results to our setting. Thus, all of the lower bounds apply equally well to the "classical special case" of shadow tomography, where we are trying to learn properties of a probability distribution $\mathcal{D}$ given independent samples from $\mathcal{D}$, and none of them yet say anything specific to quantum mechanics.

1.6 Techniques

Relating Gentleness to DP. In the proof of our main result—i.e., the connection between gentleness and differential privacy—the easy direction is that gentleness implies DP. This direction produces only constant loss in parameters, and does not even have much to do with quantum mechanics. We consider the contrapositive: if a measurement M is *not* DP, then there are two neighboring states, call them ρ and σ, as well as a measurement outcome y, such that $\Pr[M(\rho) \text{ outputs } y]$ and $\Pr[M(\sigma) \text{ outputs } y]$ differ by a large multiplicative factor. But in that case, we can study what happens if we apply M to the equal mixture $\frac{\rho+\sigma}{2}$, and then condition on outcome y. Here we can use Bayes' theorem to show that the post-measurement state will *not* be close to $\frac{\rho+\sigma}{2}$—intuitively, because it will have "more ρ than σ" or vice versa. Therefore M is not

gentle on product states (for if ρ and σ are neighbors and are themselves product states, then $\frac{\rho+\sigma}{2}$ is a product state).

The harder direction is to show that ε-DP implies $O(\varepsilon\sqrt{n})$-gentleness (for product states, and at least for a restricted class of measurements). We work up to this result in a sequence of steps. The first is to prove a purely classical analogue: namely, any ε-DP classical algorithm is $2\varepsilon\sqrt{n}$-gentle on product *distributions*, $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$—and indeed, the posterior distribution $\mathcal{D}_y$, conditioned on some output y, has KL-divergence at most $2\varepsilon^2 n$ from $\mathcal{D}$. While this step has echoes in earlier work on adaptive data analysis [19, 18, 38] (see Section 1.7), we provide our own proof for completeness. Our proof uses the ε-DP property of A, together with the fact that $\mathcal{D}$ is a product distribution, to show that, if we reveal a sample from $\mathcal{D}_y$ a single register at a time, from the 1^{st} to the n^{th}, then the expected KL-divergence from $\mathcal{D}$ increases by at most $2\varepsilon^2$ per register, and is therefore at most $2\varepsilon^2 n$ overall.

The second step is to prove an analogous result if the classical algorithm A is applied, not to a sample from the distribution $\mathcal{D}$, but in superposition to each component of the "QSampling" state

$$|\psi\rangle := \sum_x \sqrt{\Pr_{\mathcal{D}}[x]}\,|x\rangle\,.$$

To prove this, we let $|\psi_y\rangle$ be the post-measurement state conditioned on outcome y, and then upper-bound the trace distance,

$$\left\||\psi\rangle\langle\psi| - |\psi_y\rangle\langle\psi_y|\right\|_{\text{tr}} = \sqrt{1 - |\langle\psi|\psi_y\rangle|^2},$$

in terms of the square root of the KL-divergence between $\mathcal{D}_y$ and $\mathcal{D}$, which we previously showed was $O(\varepsilon\sqrt{n})$. (To do that, in turn, we use the *Hellinger distance* between $\mathcal{D}_y$ and $\mathcal{D}$ as an intermediate measure.)

The last step is to generalize from algorithms A that act separately on each computational basis state to measurements M that can apply a separate POVM to each register, and also from pure product states to mixed product states. We achieve these generalizations using standard manipulations in quantum information. We expect that further generalizations are possible with more work.

Shadow Tomography. The analysis of Quantum Private Multiplicative Weights (QPMW), our new online, gentle procedure for shadow tomography, is our technically most demanding result. The QPMW procedure itself is relatively simple,[12] and is directly inspired by an analogous procedure from classical differential privacy, the so-called Private Multiplicative Weights (PMW) algorithm of Hardt and Rothblum [27] from 2010.

Given a database $X \in [d]^n$, of n records $x_1, \ldots, x_n$ drawn independently from some underlying probability distribution $\mathcal{D}$, the goal of PMW is to answer an enormous number of statistical queries about $\mathcal{D}$, possibly as many as $\exp(n)$ of them, in a way that preserves the overall differential privacy of X. Here the queries need to be answered one by one, as they arrive, and could be chosen by an adaptive adversary.

[12] Indeed, QPMW is arguably simpler than previous shadow tomography procedures, especially because it completely avoids the use of the so-called Quantum OR Bound of Harrow, Lin, and Montanaro [31]. QPMW could, in fact, be used to give an independent proof of the OR Bound, one where the procedure would moreover be gentle (albeit, possibly with worse sample complexity).

PMW achieves this by maintaining, at all times, a current hypothesis $\mathcal{H}$ about $\mathcal{D}$. Whenever a new query arrives, the first thing PMW does is to check whether $\mathcal{H}$ and X lead to approximately the same answer for that query. If the answers are equal to within some threshold, then PMW simply answers the query using $\mathcal{H}$, without looking further at X. Only if $\mathcal{H}$ and X disagree substantially does PMW query X a second time—both to learn the correct answer to the current query, and to use that answer to update the hypothesis $\mathcal{H}$. For *both* types of queries, PMW uses the standard DP trick of adding a small amount of Laplace noise to any statistics gathered from X, before using those statistics for anything else.

It's clear, by construction, that this strange two-pronged approach will always return approximately correct answers, with high probability. But why does it help in preserving privacy? The privacy analysis depends on proving three facts:

(1) Each query leads to only a negligible loss in privacy (say, $\sim 1/\exp(n)$), *unless* it has an appreciably large probability of triggering an update.

(2) Even when an update *is* triggered, the loss in privacy is still modest, say $\sim 1/\sqrt{n}$.

(3) The number of updates is always extremely small, say $O(\log d)$. This is true for "the usual multiplicative weights reasons."

Once one understands the connection between privacy and gentleness, it's natural to wonder whether a quantum analogue of PMW might let one apply a huge sequence of measurements $E_1 \ldots, E_m$, one at a time, to a small collection of identical quantum states $\rho^{\otimes n}$ (where, say, $n \leq (\log m)^{O(1)}$), in a way that yields accurate estimates of $\Pr[E_i(\rho) \text{ accepts}]$ for every i, without destroying the states in the process or even damaging them too much. This, of course, is precisely the problem of (gentle, online) shadow tomography.

In Section 6, we prove that indeed this is possible. Our QPMW algorithm is just the "obvious" quantum generalization of PMW. That is, QPMW at all times maintains a current hypothesis, σ, about the unknown quantum state ρ. Initially σ is the maximally mixed state $\mathbb{I}/d$. Whenever a new measurement E_t arrives, QPMW first checks whether

$$\Pr[E_t(\sigma) \text{ accepts}] \approx \Pr[E_t(\rho) \text{ accepts}],$$

with the check being done using a thresholded version of the Laplace noise measurement from Section 1.1. If the answer is yes, then QPMW simply returns $\Pr[E_t(\sigma) \text{ accepts}]$ as its estimate for $\Pr[E_t(\rho) \text{ accepts}]$, without measuring the actual quantum states any further. Only if the answer is no does QPMW measure a second time—both to learn an accurate estimate for $\Pr[E_t(\rho) \text{ accepts}]$, and to use that estimate to update its hypothesis σ. This second measurement also involves the deliberate addition of Laplace noise.

Intuitively, the reason why we might expect this to work is that each round of PMW leaks very little privacy—and by our central connection between DP and gentleness, that suggests that we can implement each round of QPMW in a way that damages the states very little. However, formalizing this requires substantial new ideas, which are not contained in the classical analyses of PMW.

Of course, if we had a sufficiently general theorem about privacy implying gentleness, then perhaps everything we needed would follow immediately from that theorem, combined with the privacy of PMW. However, our existing implication—applying, as it does, only to product measurements

on product states, and saying nothing about adaptively chosen *sequences* of measurements—will force us to work harder.

The core difficulty concerns what, before, we called step (1) in the analysis of PMW: namely, the connection between loss in privacy and the probability of triggering an update. We note that while, by construction, the answer in each round is close to the answer on the current state in the registers, we need the answer to be accurate with respect to the *original* state ρ. The algorithm's gentleness plays an essential role in proving accuracy: it's only because of gentleness that we know that the state in the registers hasn't been corrupted, and that the algorithm's answers are accurate with respect to the original state. We further note that, since we want to handle many measurements, and the damage from these measurements will accumulate, we truly need to show that the overwhelming majority of measurements result in only negligible damage.

The original analysis [27] conditioned on so-called "borderline rounds," which are rounds that have a reasonable probability of triggering an update, and argued that the privacy loss in other rounds was zero. In the quantum setting, however, this is a non-starter: so long as there is some probability of an update, the damage is never zero. Instead, we show how to bound the damage each no-update round would cause to the original state as a function of the probability that it could have triggered an update. Thus, rounds that are likely to trigger an update (of which there are few) can cause damage, but rounds that are unlikely to trigger an update (of which there are many) each cause very little damage once we condition on "no update." Since the number of updates is bounded, this is a promising start. Bounding the damage as a function of the probability of an update requires a delicate analysis, leveraging the differential privacy of the Laplace measurement and the fact that we have a product state in the registers, which induces a Gaussian distribution on the answers before noise is added to each measurement (see Claim 41).

In the classical setting, once we bound the privacy loss per round, we can apply composition theorems to bound the loss across rounds. Crucially, this composition maintains *multiplicative* guarantees on the closeness of probabilities. But damage to quantum states (in the trace distance metric, for example) is additive, not multiplicative. Indeed, even if the amplitudes in a quantum state $|\psi\rangle$ were to change by only small multiplicative amounts, that could easily turn into an additive change when we rotate $|\psi\rangle$ to a different basis—a phenomenon with no classical analog. So once $|\psi\rangle$ becomes even slightly corrupted, why doesn't this sever the multiplicative connection between damage and the probability of an update—thereby preventing the necessary updates from happening, and allowing $|\psi\rangle$ to become corrupted even further, and so forth, until inaccurate answers are returned?

We address these worries using several tools. The first is a "Damage Lemma," Lemma 17, which tightly connects the probability of an update being triggered in the "real" world, where the state $\rho^{\otimes n}$ is slightly damaged by each measurement round, to its probability of being triggered in the "ideal" world, where the algorithm gets a fresh copy of $\rho^{\otimes n}$ at each round. This lemma is quite general and might find other uses. With this lemma in place, we divide the execution of the QPMW algorithm into epochs, where each epoch has a constant probability of triggering an update. By the connection between damage and update probabilities, this means that the sum of the damage incurred by an "ideal" execution would be bounded, and by the Damage Lemma the total damage in the "real" execution remains bounded as well. Since, moreover, each epoch triggers an update with constant probability, and the number of updates is bounded, the number of epochs will be bounded too. This gives us a bound on the total damage to the state, and is crucial both for proving gentleness *and* for proving accuracy.

Other Results. The paper's other results are proved using a variety of techniques. In Appendix 10, for example, we show that any measurement that's 0-DP on product states (i.e., accepts all product states with the same probability) is actually 0-DP on *all* states, and hence trivial. Though simple, this result makes essential use of the fact that the separable mixed states have positive density within the set of all mixed states, and would be *false* if amplitudes were reals rather than complex numbers. Since most results in quantum information are insensitive to the distinction between real and complex quantum mechanics, it's noteworthy to find an exception.

To prove, in Appendix 13, a weak form of composition for quantum DP algorithms, we use the same "Damage Lemma" (Lemma 17) that we used for the analysis of QPMW. In that appendix, however, we also construct an example, involving DP measurements in two incompatible bases, that shows why any composition theorem for quantum DP will come with caveats that weren't needed classically.

To prove, in Section 5, that our "DP implies gentleness" implications are asymptotically optimal, we use the Laplace noise measurement L_σ as a separating example. When $\sigma = \Theta(n)$, we get a measurement that's $O\left(\frac{1}{n}\right)$-DP, but not α-gentle on arbitrary states for any $\alpha = o(1)$. When $\sigma = \Theta(\sqrt{n})$, we get a measurement that's $O(\frac{1}{\sqrt{n}})$-DP, but not α-gentle on product states for any $\alpha = o(1)$.

1.7 Related Work

To our knowledge, this paper is the first to make the connection between gentle measurement of quantum states and differential privacy. Nevertheless, there were two previous papers that tried to combine quantum information and differential privacy in other ways; there was a previous study of gentle tomography; and there was a celebrated (purely classical) connection between differential privacy and so-called *adaptive data analysis*, which in some ways foreshadowed our connection between DP and gentle measurement.

Quantum information and DP. Senekane et al. [39] discuss first applying a classical DP algorithm to classical data, and then encoding the output into a quantum state for use in a quantum machine learning algorithm. Naturally this composition preserves DP, but the DP and quantum aspects don't seem to interact much.

Zhou and Ying [46] define and study an interesting notion of "quantum DP," which however is very different from ours. Given an algorithm A that takes a quantum state as input and produces another quantum state as output, they define A to be $(\Delta, \varepsilon, \delta)$-DP if for all states ρ, σ with trace distance at most Δ, and all 2-outcome measurements M,

$$\Pr[M(A(\rho)) \text{ accepts}] \leq e^{\varepsilon} \Pr[M(A(\sigma)) \text{ accepts}] + \delta.$$

In other words: unlike us, Zhou and Ying don't consider A's behavior on two databases that differ in a single entry (but which could have arbitrarily large trace distance)—only on two states that are actually close *as quantum states*. For them, essentially, a DP algorithm is a quantum channel that converts "mere" closeness in trace distance into a stronger, multiplicative kind of closeness between quantum states. Zhou and Ying's main results are that

(1) the standard depolarizing and amplitude-damping channels (i.e., just adding noise to a quantum state, like in the simplest models of decoherence) are DP in their sense, and

(2) their notion of quantum DP satisfies many composition theorems, including advanced composition.

These results are interesting and non-obvious, but only tangentially related to what we do.

Gentle tomography. Bennett, Harrow, and Lloyd [12] studied the task of "gentle quantum state tomography"—that is, recovering a full description of a quantum state ρ from identical copies $\rho^{\otimes n}$, without appreciably damaging the $\rho^{\otimes n}$'s. Their notion of "gentleness" was very similar to ours. To achieve the task, they gave a protocol that, like many of our protocols, deliberately adds noise to the measurement outcomes before returning them (although they used a randomized binning strategy rather than Laplace noise). They did not make a connection to differential privacy, and also did not consider shadow tomography, or any other tasks besides full tomography.

DP and adaptive data analysis. Perhaps the work that most clearly anticipated ours, at a technical level, had nothing to do with quantum information at all. Dwork et al. [19] studied the problem of *adaptive data analysis*: given a dataset, drawn i.i.d. from an underlying distribution, the goal is accurately to answer a long sequence of adaptively chosen statistical queries or analyses. Each query can be chosen as a function of the answers to all previous queries. Accuracy is measured with respect to the underlying distribution, rather than the specific dataset drawn, and the goal is to avoid overfitting. A sequence of works [19, 18, 10] showed that differentially private mechanisms are particularly well-suited to this application, and can be used to guarantee adaptive accuracy automatically.

Let X be the dataset, with n entries drawn i.i.d. from a distribution $\mathcal{D}$. A priori, before any queries are answered, an observer's view of the dataset X is that it is a draw from the distribution $\mathcal{D}^n$. As queries are answered, this view might change. One way to prevent overfitting is to guarantee that the query answers do not change the observer's view much: i.e., that the a-posteriori view of X's distribution, conditioned on the observed answers, is almost unchanged. This can be interpreted as "classical gentleness." At a technical level, our results use the fact that in the above scenario, if we run a classical DP algorithm A on the database X, then conditioning on $A(X)$ outputting any particular value y results in a bounded change to the prior (see Lemma 30). We note that a similar result follows from the work of Dwork et al. [18] and Rogers et al. [38] (their results are phrased in terms of the so-called "max information").

While there are technical and conceptual connections, the setting of quantum measurement or shadow tomography (even without gentleness) presents altogether different challenges from the adaptive data analysis setting. Most notably, as we discussed in Section 1.1, running an algorithm on a quantum state can collapse the state. This is a physical phenomenon, not just a change in a particular observer's prior and posterior views as was the case classically. In particular, quantum measurements that collapse the state cannot be forgotten or undone. Restricting our attention to computing the average of two-outcome measurements over n registers, this difference is best illustrated by the fact that, in the quantum setting, computing accurate answers to a large collection of *non-adaptive* measurements is already a challenging task (even without requiring gentleness). In the classical setting, on the other hand, if the measurements are specified non-adaptively then the naïve algorithm that simply outputs the empirical mean for each measurement performs quite well; the only challenge is answering an adaptively specified sequence of measurements.

2 Preliminaries

2.1 Classical Probability Theory

Given two probability distributions $\mathcal{D} = (p_x)_x$ and $\mathcal{D}' = (q_x)_x$, we'll use all three of the following measures of distance between them:

$$\left\|\mathcal{D} - \mathcal{D}'\right\| := \frac{1}{2}\sum_x |p_x - q_x| \quad \text{(trace distance)}$$

$$\mathrm{KL}\left(\mathcal{D}, \mathcal{D}'\right) := \sum_x p_x \ln \frac{p_x}{q_x} \quad \text{(Kullback-Leibler divergence)}$$

$$H^2\left(\mathcal{D}, \mathcal{D}'\right) := 1 - \sum_x \sqrt{p_x q_x} \quad \text{(squared Hellinger distance)}$$

Interestingly, Hellinger distance was invented in 1909, prior to the discovery of quantum mechanics, and is used for purely classical purposes in probability theory. But, as it involves the square roots of probabilities, it might be said to have a "secret affinity" for quantum mechanics that occasionally reveals itself, as it will in this paper.

Proposition 10 (Pinsker's Inequality) $\|\mathcal{D} - \mathcal{D}'\| \leq \sqrt{2\,\mathrm{KL}\left(\mathcal{D}, \mathcal{D}'\right)}$.

The following is less well-known, but we'll need it as well:

Proposition 11 (e.g. [37, p. 99]) $H\left(\mathcal{D}, \mathcal{D}'\right) \leq \sqrt{\mathrm{KL}\left(\mathcal{D}, \mathcal{D}'\right)}$.

2.2 Quantum Information Basics

In the following sections, we'll briefly review some standard notation and definitions from quantum information. More details can be found, for example, in Nielsen and Chuang [34].

A d-dimensional *pure state* is a unit vector in $\mathbb{C}^d$, which we write in ket notation as

$$|\psi\rangle = \sum_{i=1}^{d} \alpha_i |i\rangle .$$

Here $|1\rangle, \ldots, |d\rangle$ is an orthonormal basis for $\mathbb{C}^d$, and the α_i's are complex numbers called *amplitudes* satisfying $|\alpha_1|^2 + \cdots + |\alpha_d|^2 = 1$. The state $|\psi\rangle$ is also called a *superposition* over the basis states $|1\rangle, \ldots, |d\rangle$, which we can think of as the possible classical states of the system.[13] We also denote by $\langle\psi|$ the conjugate transpose of $|\psi\rangle$ (thus, $|\psi\rangle$ is a column vector while $\langle\psi|$ is a row vector). The unit-norm condition can then be written succinctly as $\langle\psi|\psi\rangle = 1$; and more generally, the complex inner product between $|\psi\rangle$ and $|\varphi\rangle$ can be written $\langle\psi|\varphi\rangle$.

In the special case $d = 2$, we call $|\psi\rangle$ a *qubit*, and typically label the orthonormal basis vectors by $|0\rangle$ and $|1\rangle$. It's also convenient to give standard names to the following two superpositions of $|0\rangle$ and $|1\rangle$:

$$|+\rangle := \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \qquad |-\rangle := \frac{|0\rangle - |1\rangle}{\sqrt{2}}.$$

[13] Note that any linear combination of the basis states $|1\rangle, \ldots, |d\rangle$, and not just $|1\rangle, \ldots, |d\rangle$ themselves, is called a "pure state."

The reader might be familiar with two types of operations that we can apply to pure states. First, given any unitary matrix U, we can map $|\psi\rangle$ to $U|\psi\rangle$. Second, we can *measure* $|\psi\rangle$ in the $|1\rangle, \ldots, |d\rangle$ basis. Doing so returns the outcome $|i\rangle$ with probability $|\alpha_i|^2 = |\langle\psi|i\rangle|^2$. Furthermore, the state $|\psi\rangle$ then "collapses" to $|i\rangle$.

More generally, we could measure $|\psi\rangle$ with respect to *any* orthonormal basis $|v_1\rangle, \ldots, |v_d\rangle$, which is equivalent to first applying a unitary U that maps each $|v_i\rangle$ to $|i\rangle$, then measuring in the $|1\rangle, \ldots, |d\rangle$ basis, and finally applying $U^\dagger$, where $\dagger$ denotes conjugate transpose. This returns the outcome $|v_i\rangle$ with probability $|\langle\psi|v_i\rangle|^2$, whereupon the state collapses to $|v_i\rangle$. A measurement of this type is called a *projective* measurement.

2.3 Mixed States, Superoperators, Quantum Operations, and POVMs

In general, we may have ordinary probabilistic uncertainty about which quantum superposition we have. This leads us to *mixed states*, the most general kind of state in quantum mechanics. Formally, a d-dimensional mixed state ρ is a $d \times d$ positive semidefinite matrix that satisfies $\mathrm{Tr}(\rho) = 1$. Equivalently, ρ is a convex combination of outer products of pure states with themselves (without loss of generality, at most d pure states):

$$\rho = \sum_{i=1}^{d} p_i |\psi_i\rangle\langle\psi_i|,$$

where $p_1, \ldots, p_d \geq 0$ and $p_1 + \cdots + p_d = 1$. This can be interpreted as a probability distribution wherein each $|\psi_i\rangle$ occurs with probability p_i, though note that different distributions can give rise to the same ρ. In the special case where $\rho = |\psi\rangle\langle\psi|$ has rank 1, it represents a pure state (i.e., a superposition). In the special case where ρ is diagonal, it represents a classical probability distribution over $|1\rangle, \ldots, |d\rangle$. The *maximally mixed state*, $\frac{\mathbb{I}}{d}$ where $\mathbb{I}$ is the identity matrix, corresponds to the uniform distribution over $|1\rangle, \ldots, |d\rangle$, and has the unique property of being unaffected by unitary transformations.

We can restate the basic rules of quantum mechanics in terms of mixed states, as follows. First, a unitary transformation U maps ρ to $U\rho U^\dagger$. Second, a measurement of ρ in the $|1\rangle, \ldots, |d\rangle$ basis returns the outcome $|i\rangle$ with probability $\rho_{ii} = \langle i|\rho|i\rangle$, whereupon ρ collapses to $|i\rangle\langle i|$. Likewise, a measurement in the $|v_1\rangle, \ldots, |v_d\rangle$ basis returns $|v_i\rangle$ with probability $\langle v_i|\rho|v_i\rangle$, whereupon ρ collapses to $|v_i\rangle\langle v_i|$.

More generally, a *superoperator* S, the most general (norm-preserving) mapping from mixed states to mixed states allowed by quantum mechanics, maps ρ to the mixed state

$$S(\rho) = \sum_{i=1}^{k} B_i \rho B_i^\dagger,$$

where $B_1, \ldots, B_k$ can be any matrices satisfying

$$\sum_{i=1}^{k} B_i^\dagger B_i = \mathbb{I}.$$

Here ρ and $S(\rho)$ do not even need to have the same dimension. Superoperators encompass unitary transformations, measurements, and other interactions with an external environment in a single formalism.

Even more generally still, if we have S as above where $B_1, \ldots, B_k$ only satisfy

$$\sum_{i=1}^{k} B_i^\dagger B_i \preceq \mathbb{I},$$

then we call S a *quantum operation*.[14] If S is a quantum operation, then $S(\rho)$ is Hermitian and positive semidefinite, but it might not be a normalized mixed state, because its trace might be less than 1. Quantum operations are useful for capturing the effects of superoperators when we additionally *condition* on some event happening (e.g., a measurement outcome being "accept"). The event's probability is then $\mathrm{Tr}(S(\rho))$, and the final mixed state conditioned on the event is $\frac{S(\rho)}{\mathrm{Tr}(S(\rho))}$.

Quantum operations act linearly on mixed states, in the sense that

$$S(p\rho + q\sigma) = pS(\rho) + qS(\sigma).$$

Although any measurement can be represented by a superoperator, when discussing measurements it's convenient to use a related formalism called "POVMs" (Positive Operator Valued Measures). POVMs capture all measurements allowed by quantum mechanics, including those whose implementations might involve ancillary systems besides the ones being measured. In this formalism, a *measurement* M is given by a list of $d \times d$ positive semidefinite matrices $E_1, \ldots, E_k$, which satisfy $E_1 + \cdots + E_k = \mathbb{I}$. The rule is:

$$\Pr[M(\rho) \text{ returns outcome } i] = \mathrm{Tr}(E_i \rho).$$

Importantly, specifying the E_i's doesn't uniquely determine the post-measurement states (i.e., what happens to ρ if the outcome is i). Thus, by an *implementation* of the measurement M, in this paper we'll mean a list of $d \times d$ matrices $B_1, \ldots, B_k$, which satisfy $B_i^\dagger B_i = E_i$. For a given implementation, if the measurement outcome is i, then the post-measurement state is

$$\frac{B_i \rho B_i^\dagger}{\mathrm{Tr}\left(B_i \rho B_i^\dagger\right)}.$$

Note that the mapping

$$\rho \rightarrow \sum_{i=1}^{k} B_i \rho B_i^\dagger$$

is a superoperator, that each individual mapping $\rho \rightarrow B_i \rho B_i^\dagger$ is a quantum operation, and that $\mathrm{Tr}\left(B_i \rho B_i^\dagger\right) = \mathrm{Tr}(E_i \rho)$ is the probability of outcome i.

In the special case of two-outcome POVMs (E_1, E_2), we'll sometimes identify the POVM itself with the "accept" outcome E_1, treating the "reject" outcome $E_2 = \mathbb{I} - E_1$ as implied.

[14] In the literature, these are also called "non-trace-increasing completely positive maps."

2.4 Separable and Entangled

A pure state $|\psi\rangle$ on n registers is called a *product state* if it can be written as a tensor product,

$$|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle .$$

Any pure state that cannot be so written is called *entangled*. A famous example of an entangled pure state is the *Bell pair*, $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$.

A mixed state ρ is likewise called a product state if it can be written as a tensor product

$$\rho = \rho_1 \otimes \cdots \otimes \rho_n.$$

Also, ρ is called *separable* if it can be written as a convex combination of product states, and *entangled* if it can't be. Unlike a pure state, a mixed state can be separable but non-product, meaning that it has classical correlations but no entanglement, as with the example $\frac{|00\rangle\langle 00|+|11\rangle\langle 11|}{2}$ (i.e., 00 and 11 with equal probabilities).

A measurement M on an n-register state is called *product* if there exist POVMs $M_1, \ldots, M_n$ such that M can be implemented as follows:

- For each $i \in [n]$, apply M_i to the i^{th} register.
- Return some function of the n classical measurement outcomes, possibly together with auxiliary randomness.

In the special case where $M_1, \ldots, M_n$ are all projective measurements, we call M a *product-of-projectives*.

More generally, we call M *mixture-of-products* if the POVMs $M_1, \ldots, M_n$ can be chosen randomly, from some correlated probability distribution, in advance of applying them.

More generally still, we call M *LOCC*—the acronym stands for Local Operations and Classical Communication—if M can be implemented by applying a POVM to some register $i_1 \in [n]$, then (depending on the outcome) applying another POVM to some register $i_2 \in [n]$, and so on, then finally returning some function of the classical measurement outcomes, possibly together with auxiliary randomness. Here we allow any finite, adaptively chosen sequence of POVMs, which could include repeated POVMs applied to the same register.

Let us stress that, even if a measurement happens to be product, or mixture-of-products, or LOCC, if we want to implement the measurement *gently*, we might need to apply a quantum circuit that acts on all n registers coherently. This is because, if we measure the registers separately, we'll generate *garbage*—i.e., information about the state besides the final measurement outcome—that might destroy gentleness. Only if we've taken care to do everything in coherent superposition, simulating the "measurements" on the individual registers (and the computations on the outcomes of those measurements) using ancilla qubits, can we later uncompute the garbage. This is likely to be a significant challenge for experimental implementation of gentle measurements like the ones discussed in this paper, since coherent measurements across n registers are much harder than incoherent ones to realize in practice. On the other hand, this issue makes no difference for DP, since even if the garbage isn't uncomputed, it need not be revealed to the end user.[15]

[15] Or to say it another way, the definition of quantum DP talks only about the probabilities of outcomes, not about the post-measurement states.

2.5 Distance Between Quantum States

Given a Hermitian matrix A, its *trace norm* is defined as

$$\|A\|_{\text{tr}} := \frac{1}{2}\sum_{i=1}^{d} |\lambda_i| ,$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of A. In particular, given two mixed states ρ and σ, their *trace distance* is defined as $\|\rho - \sigma\|_{\text{tr}}$.

Trace distance is a metric on mixed states—i.e., it's reflexive, symmetric, and satisfies the triangle inequality. It's equal to

$$\max_M \left(\Pr\left[M\left(\rho\right) \text{ accepts}\right] - \Pr\left[M\left(\sigma\right) \text{ accepts}\right]\right),$$

where the maximum is taken over all possible two-outcome measurements M. As such, trace distance generalizes the total variation distance between classical probability distributions, reducing to the latter when ρ and σ are both diagonal matrices.

We'll find the following facts useful.

Proposition 12 (Contractivity of Trace Norm [34, p. 406]) *Let S be any quantum operation, and let A be a Hermitian matrix. Then*

$$\|S\left(A\right)\|_{\text{tr}} \leq \|A\|_{\text{tr}} .$$

So in particular, for any two mixed states ρ and σ and any quantum operation S, we have

$$\|S\left(\rho\right) - S\left(\sigma\right)\|_{\text{tr}} = \|S\left(\rho - \sigma\right)\|_{\text{tr}} \leq \|\rho - \sigma\|_{\text{tr}} .$$

As an especially useful example, a superoperator that "traces out" (discards) part of its input state can never increase trace distance.

Proposition 13 (Convexity of Trace Norm) *For all Hermitian matrices A, B and $p, q \geq 0$,*

$$\|pA + qB\|_{\text{tr}} \leq p\|A\|_{\text{tr}} + q\|B\|_{\text{tr}} .$$

The triangle inequality for trace distance is just a special case of the above. As another useful special case, for all mixed states $\rho, \sigma, \rho', \sigma'$ and probabilities p,

$$\left\|\left(p\rho + (1-p)\,\sigma\right) - \left(p\rho' + (1-p)\,\sigma'\right)\right\|_{\text{tr}} \leq p\left\|\rho - \rho'\right\|_{\text{tr}} + (1-p)\left\|\sigma - \sigma'\right\|_{\text{tr}} .$$

Finally, trace distance $\|\rho - \sigma\|_{\text{tr}}$ takes an especially simple form if $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$ are both pure states.

Proposition 14 *For all $|\psi\rangle, |\phi\rangle$,*

$$\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_{\text{tr}} = \sqrt{1 - |\langle\psi|\phi\rangle|^2}.$$

2.6 Additivity of Damage

In this section, we prove the extremely useful fact that, if we apply quantum operations to a quantum state ρ in succession, then we can bound the total damage caused to ρ in trace distance by the sum of the damages that each operation *would* cause were it applied to ρ individually. This fact is related to the so-called "Quantum Union Bound" (see [3, 45]), but it's both simpler to state and easier to prove.

Lemma 15 *Let ρ be a mixed state, and let S be any quantum operation. Suppose $\|\rho' - \rho\|_{\text{tr}} \leq \varepsilon$, and let $\|S(\rho) - \rho\|_{\text{tr}} \leq \delta$. Then $\|S(\rho') - \rho\|_{\text{tr}} \leq \varepsilon + \delta$.*

Proof. We have

$$\begin{aligned}\left\|S\left(\rho'\right) - \rho\right\|_{\text{tr}} &= \left\|S\left(\rho + \left(\rho' - \rho\right)\right) - \rho\right\|_{\text{tr}} \\ &= \left\|S\left(\rho' - \rho\right) + \left(S\left(\rho\right) - \rho\right)\right\|_{\text{tr}} \\ &\leq \left\|\rho' - \rho\right\|_{\text{tr}} + \left\|S\left(\rho\right) - \rho\right\|_{\text{tr}} \\ &\leq \varepsilon + \delta.\end{aligned}$$

Here the second line used the linearity of quantum operations, and the third used the triangle inequality for trace distance as well as Proposition 12 (i.e., the fact that applying a quantum operation can never increase the trace norm). ■

Lemma 15 has the following immediate corollary.

Corollary 16 *Let ρ be a mixed state and let $S_1, \ldots, S_m$ be quantum operations. Suppose that for all i, we have*

$$\left\|S_i\left(\rho\right) - \rho\right\|_{\text{tr}} \leq \varepsilon_i.$$

Then

$$\left\|S_m\left(S_{m-1}\left(\cdots S_1\left(\rho\right)\right)\right) - \rho\right\|_{\text{tr}} \leq \varepsilon_1 + \cdots + \varepsilon_m.$$

Proof. Suppose by induction on m that

$$\left\|S_{m-1}\left(S_{m-2}\left(\cdots S_1\left(\rho\right)\right)\right) - \rho\right\|_{\text{tr}} \leq \varepsilon_1 + \cdots + \varepsilon_{m-1}.$$

Then

$$\begin{aligned}\left\|S_m\left(S_{m-1}\left(\cdots S_1\left(\rho\right)\right)\right) - \rho\right\|_{\text{tr}} &\leq \left\|S_{m-1}\left(S_{m-2}\left(\cdots S_1\left(\rho\right)\right)\right) - \rho\right\|_{\text{tr}} + \left\|S_m\left(\rho\right) - \rho\right\|_{\text{tr}} \\ &\leq \varepsilon_1 + \cdots + \varepsilon_m.\end{aligned}$$

by Lemma 15. ■

Corollary 16 is the reason why "gentleness composes": that is, applying an α_1-gentle measurement to a state ρ, followed by an α_2-gentle measurement, yields an overall $(\alpha_1 + \alpha_2)$-gentle measurement. By contrast, it's not clear to what extent DP composes in the quantum setting, because of the interaction between the DP requirement and damage to the state. For more about this issue see Appendix 13.

Note that, by simply specializing Corollary 16 to diagonal ρ and classical operations $S_1, \ldots, S_m$, we obtain an analogous statement for classical variation distance.

In Section 6, when we analyze our shadow tomography protocol, we'll also need a lemma that upper-bounds the damage caused by a sequence of measurements *conditional on the measurements all accepting*—or equivalently, by a sequence of quantum operations where we normalize the final result. Fortunately, the formalism of quantum operations and trace norm can accommodate this case as well.

Lemma 17 (Damage Lemma) *Let ρ be a mixed state. For all $i \in [m]$, let S_i be a quantum operation, which "accepts" a state σ with probability $\operatorname{Tr}(S_i(\sigma)) > 0$, and yields the post-measurement state $\frac{S_i(\sigma)}{\operatorname{Tr}(S_i(\sigma))}$ when it does. Suppose that for all $i \in [m]$, we have*

$$\left\| \frac{S_i(\rho)}{\operatorname{Tr}(S_i(\rho))} - \rho \right\|_{\text{tr}} \leq \varepsilon_i.$$

Let $p_i := \operatorname{Tr}(S_i(\rho))$ be the probability that S_i accepts the "ideal" state ρ, and let

$$q_i := \frac{\operatorname{Tr}(S_i(\cdots S_1(\rho)))}{\operatorname{Tr}(S_{i-1}(\cdots S_1(\rho)))}$$

be the probability that S_i accepts the state that it actually receives, if $S_1, \ldots, S_{i-1}$ are first applied to ρ and if we condition on their accepting. Given any subset $T \subseteq [m]$, let

$$p_T := \prod_{i \in T} p_i, \qquad q_T := \prod_{i \in T} q_i.$$

Then for all T,

$$|p_T - q_T| \leq \frac{\varepsilon_1 + \cdots + \varepsilon_m}{q_{[m] \setminus T}}.$$

Also,

$$\left\| \frac{S_m(\cdots S_1(\rho))}{\operatorname{Tr}(S_m(\cdots S_1(\rho)))} - \rho \right\|_{\text{tr}} \leq \frac{2}{q_{[m]}} (\varepsilon_1 + \cdots + \varepsilon_m).$$

Proof. For all $i \in [m]$, let

$$E_i := \frac{S_i(\rho)}{\operatorname{Tr}(S_i(\rho))} - \rho.$$

Then by hypothesis, $\|E_i\|_{\text{tr}} \leq \varepsilon_i$. Also,

$$S_i(\rho) = p_i(\rho + E_i).$$

We can now write:

$$\begin{aligned}
S_1(\rho) &= p_1(\rho + E_1), \\
S_2(S_1(\rho)) &= p_1(S_2(\rho) + S_2(E_1)) \\
&= p_1(p_2(\rho + E_2) + S_2(E_1)), \\
S_3(S_2(S_1(\rho))) &= p_1(p_2(S_3(\rho) + S_3(E_2)) + S_3(S_2(E_1))) \\
&= p_1(p_2(p_3(\rho + E_3) + S_3(E_2)) + S_3(S_2(E_1))),
\end{aligned}$$

and so on until

$$\begin{aligned} S_m\left(\cdots S_1\left(\rho\right)\right) = p_{[m]}\rho + p_{[m]}E_m + p_{[m-1]}S_m\left(E_{m-1}\right) \\ + p_{[m-2]}S_m\left(S_{m-1}\left(E_{m-2}\right)\right) + \cdots + p_{[1]}S_m\left(\cdots S_2\left(E_1\right)\right). \end{aligned}$$

More generally, suppose we define

$$S'_i\left(\sigma\right) := \begin{cases} S_i\left(\sigma\right) & \text{if } i \in T \\ \frac{S_i(\sigma)}{q_i} & \text{otherwise,} \end{cases}$$

so that

$$\mathrm{Tr}\left(S'_m\left(\cdots S'_1\left(\rho\right)\right)\right) = \frac{q_1 \cdots q_m}{\prod_{i \notin T} q_i} = q_T$$

is just the probability that S_i accepts for all $i \in T$, if $S_1, \ldots, S_m$ are applied in sequence. Then repeating the manipulations above gives us the following modified equation, in which all the products of p_i's are restricted to range only over $i \in T$:

$$\begin{aligned} S'_m\left(\cdots S'_1\left(\rho\right)\right) = p_T\left(\rho + E_m\right) + p_{T\cap[m-1]}S'_m\left(E_{m-1}\right) \\ + p_{T\cap[m-2]}S'_m\left(S'_{m-1}\left(E_{m-2}\right)\right) + \cdots + p_{T\cap[1]}S'_m\left(\cdots S'_2\left(E_1\right)\right). \end{aligned}$$

Hence

$$\begin{aligned} \left\|S'_m\left(\cdots S'_1\left(\rho\right)\right) - p_T\rho\right\|_{\mathrm{tr}} &\leq p_T\left\|E_m\right\|_{\mathrm{tr}} + p_{T\cap[m-1]}\left\|S'_m\left(E_{m-1}\right)\right\|_{\mathrm{tr}} + \cdots + p_{T\cap[1]}\left\|S'_m\left(\cdots S'_2\left(E_1\right)\right)\right\|_{\mathrm{tr}} \\ &\leq \left\|E_m\right\|_{\mathrm{tr}} + \left\|S'_m\left(E_{m-1}\right)\right\|_{\mathrm{tr}} + \cdots + \left\|S'_m\left(\cdots S'_2\left(E_1\right)\right)\right\|_{\mathrm{tr}} \\ &\leq \varepsilon_m + \frac{\varepsilon_{m-1}}{q_{\{m\}\setminus T}} + \cdots + \frac{\varepsilon_1}{q_{\{2,\ldots,m\}\setminus T}} \\ &\leq \frac{\varepsilon_1 + \cdots + \varepsilon_m}{q_{[m]\setminus T}}, \end{aligned}$$

where the second line used the fact that all the products of p_i's are upper-bounded by 1. This means that

$$\left|q_T - p_T\right| = \left|\mathrm{Tr}\left(S'_m\left(\cdots S'_1\left(\rho\right)\right)\right) - p_T\right| \leq \frac{\varepsilon_1 + \cdots + \varepsilon_m}{q_{[m]\setminus T}},$$

thereby proving the first part of the lemma.

For the second part, let us take the special case $T = [m]$. Then $q_{[m]\setminus T} = 1$, and the inequalities above reduce to

$$\begin{aligned} \left\|S_m\left(\cdots S_1\left(\rho\right)\right) - p_{[m]}\rho\right\|_{\mathrm{tr}} &\leq \varepsilon_1 + \cdots + \varepsilon_m, \\ \left|q_{[m]} - p_{[m]}\right| &\leq \varepsilon_1 + \cdots + \varepsilon_m. \end{aligned}$$

So the triangle inequality gives

$$\begin{aligned} \left\|S_m\left(\cdots S_1\left(\rho\right)\right) - q_{[m]}\rho\right\|_{\mathrm{tr}} &\leq \left\|S_m\left(\cdots S_1\left(\rho\right)\right) - p_{[m]}\rho\right\|_{\mathrm{tr}} + \left\|p_{[m]}\rho - q_{[m]}\rho\right\|_{\mathrm{tr}} \\ &\leq \varepsilon_1 + \cdots + \varepsilon_m + \left|p_{[m]} - q_{[m]}\right| \\ &\leq 2\left(\varepsilon_1 + \cdots + \varepsilon_m\right). \end{aligned}$$

Hence

$$\left\| \frac{S_m \left(\cdots S_1 \left(\rho \right) \right)}{q_{[m]}} - \rho \right\|_{\text{tr}} \leq \frac{2}{q_{[m]}} \left(\varepsilon_1 + \cdots + \varepsilon_m \right).$$

■

As we'll show in Appendix 13, Lemma 17 implies a limited sort of composition for quantum DP algorithms. Namely, we can sequentially compose k quantum DP algorithms and have the result remain accurate and DP, so long as the total damage incurred to the quantum state (in trace distance) is always small compared to the joint probability of the observed outcomes $y_1, \ldots, y_k$. We can sometimes ensure the latter property, in turn, by using our main result, the connection between DP and gentleness.

Note that we can combine Lemmas 17 and 15, to say that, even if we apply a final superoperator S_{m+1} *after* applying the quantum operations $S_1, \ldots, S_m$ and then conditioning on their results, the total damage to our initial state ρ is at most $\left\| S_{m+1} \left(\rho \right) - \rho \right\|_{\text{tr}}$ plus the damage bound from Lemma 17. (This wouldn't be true if we'd composed in the opposite order, since conditioning could amplify earlier damage to ρ by an $O\left(\frac{1}{p}\right)$ factor.) This fact will also be used in Section 6.

2.7 Pure vs. Mixed States

We now prove two propositions to show that, when considering differential privacy and gentle measurements, we can restrict attention to pure states without loss of generality; our conclusions will then automatically carry over to mixed states.

Proposition 18 *If M is ε-DP on pure product states, then M is ε-DP on mixed product states as well. Likewise, if M is ε-DP on all pure states, then M is ε-DP on all mixed states.*

Proof. Suppose we seek to maximize the ratio

$$\frac{\Pr \left[M \left(\rho \right) = y \right]}{\Pr \left[M \left(\sigma \right) = y \right]}$$

over product states $\rho = \rho_1 \otimes \cdots \otimes \rho_n$ and $\sigma = \sigma_1 \otimes \cdots \otimes \sigma_n$ that differ only on the i^{th} register. Then holding the other $n-1$ registers fixed, we're maximizing over ρ_i and minimizing over σ_i. By convexity, the maximum and minimum will both always be achieved by pure states. A second appeal to convexity then shows that the maximum ratio is also achieved when the other $n-1$ registers are set to pure states as well.

For the second part, the argument is the same, except that we simply maximize $\Pr \left[M \left(\rho \right) = y \right]$ over *all* ρ, and minimize $\Pr \left[M \left(\sigma \right) = y \right]$ over all σ. ■

Proposition 19 *If the measurement M is α-gentle on pure product states, then M is α-gentle on mixed product states as well. Likewise, if M is α-gentle on all pure states, then M is α-gentle on all mixed states.*

Proof. Fix an implementation of M; the same implementation that achieves gentleness on pure states will also achieve gentleness on mixed states.

Suppose we apply M to the product state $\rho = \rho_1 \otimes \cdots \otimes \rho_n$. As a first step, we can purify $\rho_1, \ldots, \rho_n$ to $\left| \psi_1 \right\rangle, \ldots, \left| \psi_n \right\rangle$ respectively by adding registers to them. Then M can be seen as acting on the pure state $\left| \psi \right\rangle = \left| \psi_1 \right\rangle \otimes \cdots \otimes \left| \psi_n \right\rangle$, and simply ignoring these added registers. By assumption,

after we apply M and condition on some outcome y, we're left with a post-measurement state $|\psi_y\rangle$ such that

$$\left\| |\psi_y\rangle\langle\psi_y| - |\psi\rangle\langle\psi| \right\|_{\text{tr}} \leq \alpha.$$

Likewise, let ρ_y be the post-measurement state if we apply M to ρ and then condition on outcome y. Observe that ρ_y is *also* the mixed state obtained by starting from $|\psi_y\rangle$ and then tracing out the added registers. So by Proposition 12, we have $\|\rho_y - \rho\|_{\text{tr}} \leq \alpha$ as well.

For the second part, the argument is the same, except that we purify ρ as a whole rather than $\rho_1, \ldots, \rho_n$ separately. ■

3 Basic Relations Among DP, Gentleness, and Triviality

In this section, we prove our first connection between the differential privacy and the gentleness of quantum measurements:

Theorem 20 *If a measurement M is ε-DP on all states, then M is $O(\varepsilon n)$-gentle on all states. Conversely, if M is α-gentle on all states for $\alpha \leq \frac{1}{4.01}$, then M is $O(\alpha)$-DP on all states.*

Unfortunately, Theorem 20 is weaker than it might look, since as we'll see, it relates DP to gentleness only in a regime where M is "nearly trivial." Later, we'll restrict our attention to product states, which will lead to a much more interesting connection between DP and gentleness. Nevertheless, Theorem 20 serves as an instructive warmup to our main results, and the tools used to prove it will later be reused.

Note that all the results in this section also have classical analogues—we simply need to replace "all (mixed) states" by "all probability distributions" in each definition and statement—and those classical analogues might be of interest as well.

Let's first define formally what we mean by a measurement being "nearly trivial."

Definition 21 (Triviality) *Given a set S of mixed states, a measurement M, and a parameter $\varepsilon \geq 0$, we say M is ε-**trivial on** S if for all states $\rho, \sigma \in S$, and all possible outcomes y of M, we have*

$$\Pr[M(\rho) \text{ outputs } y] \leq e^{\varepsilon} \Pr[M(\sigma) \text{ outputs } y].$$

For M to be ε-trivial, full stop, means that M is ε-trivial on the set of all states.

In particular, M is 0-trivial if and only if M's output probabilities are completely independent of ρ. Note also that ε-trivial immediately implies ε-DP. Like ε-DP (but unlike α-gentleness), the definition of ε-triviality depends only on the outcome probabilities, and not on the post-measurement states.

The following proposition gives a slightly weaker condition that already suffices for a measurement to be ε-trivial.

Proposition 22 *Given a measurement M and parameter $\varepsilon \geq 0$, suppose that for every two orthogonal pure states $|\psi\rangle$ and $|\phi\rangle$, and every possible outcome y of M, we have*

$$\Pr[M(|\psi\rangle) \text{ outputs } y] \leq e^{\varepsilon} \Pr[M(|\phi\rangle) \text{ outputs } y].$$

Then M is ε-trivial.

Proof. Let $E_1 + \cdots + E_k = \mathbb{I}$ be the POVM elements of M. Assume without loss of generality that the outcome y corresponds to the element $E = E_1$. Then by assumption,

$$\langle\psi|E|\psi\rangle \leq e^{\varepsilon} \langle\phi|E|\phi\rangle \tag{5}$$

for all orthogonal $|\psi\rangle, |\phi\rangle$. But this means that all of E's eigenvalues must be within an e^{ε} multiplicative factor of each other. So (5) holds for *all* $|\psi\rangle, |\phi\rangle$, not just all orthogonal $|\psi\rangle, |\phi\rangle$. By convexity, we then have

$$\operatorname{Tr}(E\rho) \leq e^{\varepsilon} \operatorname{Tr}(E\sigma)$$

for all ρ, σ as well. ■

Using Proposition 22, we now show that gentleness on all states implies near-triviality.

Lemma 23 *Suppose M is α-gentle on all states. Then M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-trivial—so in particular, $O(\alpha)$-trivial, provided $\alpha \leq \frac{1}{4.01}$.*

Proof. Given mixed states ρ and σ, let's first consider the special case where ρ and σ are perfectly distinguishable (that is, $\|\rho - \sigma\|_{\text{tr}} = 1$). For any outcome y, let $p := \Pr[M(\rho) \text{ outputs } y]$ and $q := \Pr[M(\sigma) \text{ outputs } y]$, and assume without loss of generality that $p \geq q$ and $p > 0$. Also, fix an α-gentle implementation of M. Let ρ_y and σ_y be the post-measurement states for ρ and σ respectively, if the outcome of applying M is y. Now consider the mixed state $\xi := \frac{\rho+\sigma}{2}$. *Its* post-measurement state is

$$\xi_y = \frac{p\rho_y + q\sigma_y}{p+q}.$$

So let $\delta := \frac{p}{p+q} - \frac{1}{2}$. Then

$$\begin{aligned}\xi - \xi_y &= \frac{1}{2}\rho + \frac{1}{2}\sigma - \frac{p}{p+q}\rho_y - \frac{q}{p+q}\sigma_y \\ &= \frac{p}{p+q}(\rho - \rho_y) + \frac{q}{p+q}(\sigma - \sigma_y) - \delta(\rho - \sigma).\end{aligned}$$

So by the triangle inequality,

$$\begin{aligned}\delta\|\rho - \sigma\|_{\text{tr}} &\leq \|\xi - \xi_y\|_{\text{tr}} + \frac{p}{p+q}\|\rho - \rho_y\|_{\text{tr}} + \frac{q}{p+q}\|\sigma - \sigma_y\|_{\text{tr}} \\ &\leq 2\alpha,\end{aligned}$$

since $\|\xi - \xi_y\|_{\text{tr}}$ and $\|\rho - \rho_y\|_{\text{tr}}$ and $\|\sigma - \sigma_y\|_{\text{tr}}$ are all at most α by our gentleness assumption. Furthermore, by assumption, $\|\rho - \sigma\|_{\text{tr}} = 1$. Thus we simply get $\delta \leq 2\alpha$. Or

$$\frac{\Pr[M(\rho) \text{ outputs } y]}{\Pr[M(\sigma) \text{ outputs } y]} = \frac{p}{q} = \frac{\frac{1}{2}+\delta}{\frac{1}{2}-\delta} \leq \frac{1+4\alpha}{1-4\alpha}.$$

And by Proposition 22, if the above holds for perfectly distinguishable states ρ, σ (so in particular, for orthogonal pure states), then it holds for all ρ, σ as well. Hence M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-trivial. ■

An immediate corollary of Lemma 23 is this:

Corollary 24 *If M is α-gentle on all states, then M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-DP on all states.*

Indeed, since the reasoning applied independently to each measurement outcome y, we get the following stronger conclusion, which will be useful when we analyze shadow tomography:

Corollary 25 *If M is (α,δ)-gentle on all states, then M is $\left(\ln\left(\frac{1+4\alpha}{1-4\alpha}\right),\delta\right)$-DP on all states.*

Notice that the central gambit in the proof of Lemma 23, namely defining $\xi := \frac{\rho+\sigma}{2}$, generally maps product states to non-product states. It turns out that this is inherent: Lemma 23 does *not* have an analogue that assumes only gentleness on product states. Or rather: if we assume only gentleness on product states, then we can deduce DP (and will do so, in Lemma 28), but will not be able to deduce triviality. And this is to be expected, since there *are* nontrivial DP algorithms, and indeed our main result (Theorem 5) shows that these algorithms lead to measurements that are gentle on product states.

We next prove a converse to Lemma 23: that near-triviality implies gentleness.

Lemma 26 (Trivial$\Longrightarrow$Gentle) *Suppose M is ε-trivial. Then M is $(e^{\varepsilon}-1)$-gentle on all states—so in particular, $O(\varepsilon)$-gentle, provided (say) $\varepsilon \leq 1$.*

Proof. Again, let $E_1+\cdots+E_k=\mathbb{I}$ be the POVM elements of M, and recall that we can use any solutions to the equations $E_i = A_i^{\dagger}A_i$ to define the possible post-measurement states after M is applied. Without loss of generality, focus on $E=E_1$ and $A=A_1$.

Since M is ε-trivial, all of E's eigenvalues must be within an e^{ε} multiplicative factor of each other. Also, since E is Hermitian, we can diagonalize it as $U^{\dagger}DU$, where U is unitary and D is a diagonal matrix of E's eigenvalues. Let's make the choice $A := U^{\dagger}\sqrt{D}U$. Then for some constant $0<c<1$, we can write $\sqrt{D}$ as $c(I+\delta V)$, where $\delta \leq 1-e^{-\varepsilon/2}$, and V is a diagonal matrix whose entries are all at most 1 in absolute value.

Let $|\psi\rangle$ be a pure state to which M is applied, and assume $\langle\psi|E|\psi\rangle > 0$. Then conditioning on outcome E leads to the post-measurement state

$$\frac{U^{\dagger}\sqrt{D}U|\psi\rangle}{\left\|U^{\dagger}\sqrt{D}U|\psi\rangle\right\|}.$$

Therefore the post-measurement state is

$$\frac{cU^{\dagger}(I+\delta V)U|\psi\rangle}{\|cU^{\dagger}(I+\delta V)U|\psi\rangle\|} = \frac{|\psi\rangle+\delta U^{\dagger}VU|\psi\rangle}{\||\psi\rangle+\delta U^{\dagger}VU|\psi\rangle\|}$$

By Proposition 14, the trace distance between this state and $|\psi\rangle$ is at most the Euclidean distance, which in turn is at most

$$\frac{1+\delta}{1-\delta}-1 \leq \frac{2-e^{-\varepsilon/2}}{e^{-\varepsilon/2}}-1 \leq e^{\varepsilon}-1.$$

Thus, we've given an implementation of M that is $(e^{\varepsilon}-1)$-gentle on pure states. By Proposition 19, this implies that M is $(e^{\varepsilon}-1)$-gentle on mixed states as well. ■

Finally, we prove that if a measurement is ε-DP *for sufficiently small* ε, then it's nearly trivial.

Proposition 27 (Sufficiently DP Is Trivial) *If M is ε-DP on all states, then M is $2\varepsilon n$-trivial on all states.*

Proof. Let ρ, σ be any mixed states on n registers. Also, let S_i be a superoperator that simply swaps out the i^{th} register for some fixed state—say the maximally mixed state $\mathbb{I}/d$, if the registers are d-dimensional.[16] Then by applying all n of the S_i's to ρ or σ, one at a time, we can map the entire input state to $\mathbb{I}/d^n$. Thus, for any output possible y of M, if we repeatedly invoke the fact that M is ε-DP, once for each S_i, we find that

$$\Pr\left[M\left(\rho\right) \text{ outputs } y\right] \leq e^{\varepsilon n} \Pr\left[M\left(\mathbb{I}/d^n\right) \text{ outputs } y\right].$$

Likewise,

$$\Pr\left[M\left(\sigma\right) \text{ outputs } y\right] \geq e^{-\varepsilon n} \Pr\left[M\left(\mathbb{I}/d^n\right) \text{ outputs } y\right].$$

Hence

$$\Pr\left[M\left(\rho\right) \text{ outputs } y\right] \leq e^{2\varepsilon n} \Pr\left[M\left(\sigma\right) \text{ outputs } y\right].$$

■

One can show, by a similar argument, that if M is ε-DP on product states, then M is $2\varepsilon n$-trivial on product states. Again, though, this is only interesting in the regime $\varepsilon \ll \frac{1}{n}$, whereas our results in Section 4 will be able to handle measurements that are ε-DP on product states for ε up to about $\frac{1}{\sqrt{n}}$.

Combining Lemma 23, Lemma 27, and Lemma 26 now completes the proof of Theorem 20.

Again, the problem with Theorem 20 is that, while it relates the privacy of a measurement M to its gentleness, it does so only as an "accidental byproduct" of showing that sufficiently private and sufficiently gentle measurements are both nearly trivial. To get a more interesting connection between privacy and gentleness, we'll need to restrict our attention to product states, as our main result (Theorem 5) does.

4 Proof of Main Result

In this section we prove Theorem 5, the two-way connection between gentleness and differential privacy on product states. Unlike Theorem 20, this connection will work even for measurements that are very far from trivial.

4.1 Gentleness Implies DP on Product States

We'll start by proving the "easy" direction: that gentleness on product states implies differential privacy on product states. For this, we can reapply Lemma 23 from the previous section.

Lemma 28 (Gentleness$\Longrightarrow$DP on Product States) *If M is α-gentle on product states, then M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-DP on product states as well—so in particular, $O\left(\alpha\right)$-DP on product states, provided $\alpha \leq \frac{1}{4.01}$. Likewise, if M is (α, δ)-gentle on product states then M is $\left(\ln\left(\frac{1+4\alpha}{1-4\alpha}\right), \delta\right)$-DP on product states.*

Proof. Let $\rho = \rho_1 \otimes \cdots \otimes \rho_n$ and $\sigma = \sigma_1 \otimes \cdots \otimes \sigma_n$ be two product states that differ only on the i^{th} register. Also, fix an implementation of M that is α-gentle on product states. Then for

[16] Or if we preferred unitary transformations, we could also achieve the same effect by (for example) applying a Haar-random unitary U to the i^{th} register, and then appealing to convexity.

any outcome y, let ρ_y and σ_y be the post-measurement states for ρ and σ respectively assuming that M returns outcome y, and let $\rho_{y,i}$ and $\sigma_{y,i}$ be the restrictions (i.e., partial traces) of ρ_y and σ_y respectively to the i^{th} register. Then by Proposition 12, together with the assumption of α-gentleness, we have

$$\|\rho_{y,i} - \rho_i\|_{\text{tr}} \leq \|\rho_y - \rho\|_{\text{tr}} \leq \alpha,$$

and likewise

$$\|\sigma_{y,i} - \sigma_i\|_{\text{tr}} \leq \|\sigma_y - \sigma\|_{\text{tr}} \leq \alpha.$$

But now we can apply Lemma 23—which implies that, if we think of M as acting on the i^{th} register only, with the other $n-1$ registers held fixed, then M must be $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-trivial. Moreover, the preceding statement holds for *all* i, and *all* settings of the other $n-1$ registers. But that's simply another way of saying that M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-DP on product states.

The last part follows simply because the argument applies for each possible output y independently. ■

This proves part (1) of Theorem 5.

Note that as α approaches $\frac{1}{4}$, the bound on the DP parameter diverges. Certainly the DP parameter needs to diverge as α approaches $\frac{1}{2}$, since (for example) measuring a single qubit in the $\{|0\rangle, |1\rangle\}$ basis, outputting the result, and then replacing the qubit by the maximally mixed state is $\frac{1}{2}$-gentle but preserves no privacy whatsoever. We leave it as an open problem to close the gap between $\frac{1}{4}$ and $\frac{1}{2}$.

4.2 DP Implies Gentleness On Product States

For the other direction, we'll proceed in stages. We'll start by providing that ε-DP implies $O\left(\varepsilon\sqrt{n}\right)$-gentleness for classical product distributions. Later we'll extend this result to the quantum setting.

We'll need a claim, originally proved in [22], that's found many uses in classical DP.

Claim 29 ([22]) *Suppose two probability distributions, $\mathcal{D} = (p_x)_{x\in[d]}$ and $\mathcal{D}' = (q_x)_{x\in[d]}$, satisfy*

$$\left|\ln \frac{p_x}{q_x}\right| \leq \varepsilon$$

for all $x \in [d]$. Then the KL-divergence,

$$\text{KL}\left(\mathcal{D}, \mathcal{D}'\right) = \sum_{x=1}^{d} p_x \ln \frac{p_x}{q_x},$$

satisfies $\text{KL}\left(\mathcal{D}, \mathcal{D}'\right) \leq 2\varepsilon^2$.

We can now prove a classical "DP implies gentleness" result. As noted previously, similar results follow from the work of Dwork et al. [18] and Rogers et al. [38] (phrased in terms of the so-called "max information"), but we provide a self-contained proof.

Lemma 30 (Classical DP$\Longrightarrow$Gentleness) *Let A be a classical ε-DP algorithm, and let $\mathcal{D}$ be a product distribution over databases X. Then for all possible outputs y of A, the posterior distribution $\mathcal{D}_y$ satisfies $\|\mathcal{D}_y - \mathcal{D}\| \leq 2\varepsilon\sqrt{n}$, and indeed the stronger bound $\text{KL}\left(\mathcal{D}_y, \mathcal{D}\right) \leq 2\varepsilon^2 n$.*

Proof. Fix any output y. We want to compare the prior distribution $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ over databases X to the posterior distribution $\mathcal{D}_y$, which is obtained by conditioning on the event $A(X) = y$. To do this, consider a process wherein we draw a database $X = (x_1, \ldots, x_n)$ from $\mathcal{D}_y$, by first drawing x_1 from the marginal distribution over the first entry conditioned on $A(X) = y$, then drawing x_2 from the marginal distribution over the second entry conditioned on $A(X) = y$ and on x_1, and so on up to x_n.

Let's call the i^{th} distribution above $\mathcal{T}_i$; note that $\mathcal{T}_i$ depends both on y and on $x_1, \ldots, x_{i-1}$. These are distributions over $[d]$, our "data universe." We claim that, for every possible value $v \in [d]$ for x_i, the log-ratio between v's probabilities under $\mathcal{T}_i$ and under $\mathcal{D}_i$ must be upper-bounded in magnitude by ε. To show this, let $W := (x_1, \ldots, x_{i-1})$ and $Z := (x_{i+1}, \ldots, x_n)$. Then

$$\begin{aligned}
\frac{\Pr_{\mathcal{T}_i}[v]}{\Pr_{\mathcal{D}_i}[v]} &= \frac{\Pr[v \mid y, W]}{\Pr[v]} \\
&= \frac{\Pr[v \mid y, W]}{\Pr[v \mid W]} \\
&= \frac{\Pr[y \mid W, v]}{\Pr[y \mid W]} \\
&= \frac{\sum_Z \Pr[Z \mid W, v] \Pr[y \mid W, v, Z]}{\sum_Z \Pr[Z \mid W] \Pr[y \mid W, Z]} \\
&= \frac{\sum_Z \Pr[Z] \Pr[y \mid W, v, Z]}{\sum_Z \Pr[Z] \Pr[y \mid W, Z]}.
\end{aligned}$$

Here the second and last lines used the assumption that $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ is a product distribution. Also, by differential privacy,

$$e^{-\varepsilon} \leq \frac{\Pr[y \mid W, v, Z]}{\Pr[y \mid W, Z]} \leq e^{\varepsilon}$$

for all y, W, v, Z. Therefore convex combinations of the above probabilities are also within an e^{ε} multiplicative factor of one another, so

$$\left| \ln \frac{\Pr_{\mathcal{T}_i}[v]}{\Pr_{\mathcal{D}_i}[v]} \right| \leq \varepsilon.$$

By Claim 29, this means that the *expected* log-ratio between $\Pr_{\mathcal{T}_i}[v]$ and $\Pr_{\mathcal{D}_i}[v]$, with respect to x_i drawn from $\mathcal{T}_i$, is upper-bounded by $2\varepsilon^2$:

$$\mathrm{KL}(\mathcal{T}_i, \mathcal{D}_i) = \mathrm{E}_{v \sim \mathcal{T}_i}\left[\ln \frac{\Pr_{\mathcal{T}_i}[v]}{\Pr_{\mathcal{D}_i}[v]} \right] \leq 2\varepsilon^2.$$

Furthermore, the expected sum of the log-ratios—i.e., the KL-divergence between $\mathcal{D}_y$ and $\mathcal{D}$ themselves—is just the sum of the expected log-ratios:

$$\mathrm{KL}(\mathcal{D}_y, \mathcal{D}) = \sum_{i=1}^{n} \mathrm{E}\left[\mathrm{KL}(\mathcal{T}_i, \mathcal{D}_i)\right] \leq 2\varepsilon^2 n,$$

where the expectations here are over the choices for the $\mathcal{T}_i$'s (which, however, are irrelevant to the upper bound). So by Pinsker's inequality (Proposition 10),

$$\|\mathcal{D}_y - \mathcal{D}\| \leq \sqrt{2\,\mathrm{KL}(\mathcal{D}_y, \mathcal{D})} \leq 2\varepsilon\sqrt{n}.$$

■

Having shown that ε-DP implies $O(\varepsilon\sqrt{n})$-gentleness for classical product distributions, we now begin the task of extending the result to quantum product states.

Lemma 31 *Suppose the measurement M is ε-DP on product states, and is a product-of-projectives (i.e., consists of a classical algorithm applied to the outcomes of nonadaptive projective measurements on the n registers). Then M is $O(\varepsilon\sqrt{n})$-gentle on product states.*

Proof. By Proposition 19, it suffices to give an implementation of M that is $O(\varepsilon\sqrt{n})$-gentle on *pure* product states. Thus, let

$$|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle = \sum_{X \in [d]^n} \alpha_X |X\rangle$$

be a pure product state on n registers. By applying suitable local unitaries, we can assume without loss of generality that M simply measures each $|\psi_i\rangle$ in the computational basis, obtaining the string $|X\rangle = |x_1\rangle \cdots |x_n\rangle$ with probability $|\alpha_X|^2$. It then outputs a sample from some probability distribution $\mathcal{M}_X$, depending on X, over the possible outputs y. We need to show how to sample from $\mathcal{M}_X$ in an $O(\varepsilon\sqrt{n})$-gentle manner.

Our implementation is as follows: first map the state $|\psi\rangle$ to

$$\sum_{X \in [d]^n : \Pr[X] > 0,\ y} \alpha_X |X\rangle \sqrt{\Pr[y|X]}\, |y\rangle .$$

Note that, as long as we do not care about computational complexity, the above mapping can always be implemented *somehow*, although implementing it efficiently requires an efficient algorithm for "QSampling" the probability distributions $\mathcal{M}_X$. Next, measure the $|y\rangle$ register in the computational basis, and condition on getting some particular result y.

Then by the rules of quantum mechanics and Bayes' rule, the state of the first register is just

$$\begin{aligned}
|\psi_y\rangle :=& \frac{\sum_{X \in [d]^n : \Pr[X] > 0} \alpha_X \sqrt{\Pr[y|X]}\, |X\rangle}{\sqrt{\sum_{X \in [d]^n : \Pr[X] > 0} |\alpha_X|^2 \Pr[y|X]}} \\
&= \frac{\sum_{X \in [d]^n : \Pr[X] > 0} \alpha_X \sqrt{\Pr[y|X]}\, |X\rangle}{\sqrt{\Pr[y]}} \\
&= \sum_{X \in [d]^n : \Pr[X] > 0} \alpha_X \sqrt{\frac{\Pr[y|X]}{\Pr[y]}}\, |X\rangle \\
&= \sum_{X \in [d]^n : \Pr[X] > 0} \alpha_X \sqrt{\frac{\Pr[X|y]}{\Pr[X]}}\, |X\rangle \\
&= \sum_{X \in [d]^n : \Pr[X] > 0} \frac{\alpha_X}{|\alpha_X|} \sqrt{\Pr[X|y]}\, |X\rangle .
\end{aligned}$$

Let $\mathcal{D}$ be the distribution over $X \in [d]^n$ defined by $\Pr_{\mathcal{D}}[X] = |\alpha_X|^2$; note that $\mathcal{D}$ is a product distribution, to which Lemma 30 applies. Also, let $\mathcal{D}_y$ be $\mathcal{D}$ conditioned on the event that M

outputs y. Then we see above that $|\psi_y\rangle$ is a pure state that precisely corresponds to $\mathcal{D}_y$—in the sense that, if we measure $|\psi_y\rangle$ in the computational basis, we'll see a sample from $\mathcal{D}_y$. The one complication is that $|\psi_y\rangle$ has an additional set of degrees of freedom, namely the unit-norm phases $\frac{\alpha_X}{|\alpha_X|}$. However, even these phases go away when we calculate the inner products $\langle\psi|\psi_y\rangle$ (which involve complex conjugates). In more detail:

$$\begin{aligned}\langle\psi|\psi_y\rangle &= \sum_{X\in[d]^n:\Pr[X]>0} \alpha_X^* \frac{\alpha_X}{|\alpha_X|}\sqrt{\Pr\left[X|y\right]} \\ &= \sum_{X\in[d]^n} |\alpha_X| \sqrt{\Pr\left[X|y\right]} \\ &= \sum_{X\in[d]^n} \sqrt{\Pr\left[X\right]\Pr\left[X|y\right]} \\ &= 1 - H^2\left(\mathcal{D},\mathcal{D}_y\right).\end{aligned}$$

Here $H^2\left(\mathcal{D},\mathcal{D}_y\right)$ is the *squared Hellinger distance* between the probability distributions $\mathcal{D}$ and $\mathcal{D}_y$ (see Section 2.1). So in particular, $\langle\psi|\psi_y\rangle$ strictly relates the distributions $\mathcal{D}$ and $\mathcal{D}_y$, and has nothing further to do with quantum mechanics.

We can now upper-bound the trace distance between $|\psi\rangle$ and $|\psi_y\rangle$—and hence, the gentleness of M on $|\psi\rangle$—by

$$\begin{aligned}\left\||\psi\rangle\langle\psi| - |\psi_y\rangle\langle\psi_y|\right\|_{\text{tr}} &= \sqrt{1-\left|\langle\psi|\psi_y\rangle\right|^2} \\ &= \sqrt{1-\left(1-H^2\left(\mathcal{D},\mathcal{D}_y\right)\right)^2} \\ &= \sqrt{2H^2\left(\mathcal{D},\mathcal{D}_y\right) - H^4\left(\mathcal{D},\mathcal{D}_y\right)} \\ &\leq \sqrt{2}H\left(\mathcal{D},\mathcal{D}_y\right) \\ &\leq \sqrt{2\,\text{KL}\left(\mathcal{D}_y,\mathcal{D}\right)} \\ &\leq 2\sqrt{2}\cdot\varepsilon\sqrt{n}.\end{aligned}$$

Here the first line used Proposition 14, the second-to-last line used Proposition 11, and the last line used Lemma 30. ■

Note that, if we'd upper-bounded the Hellinger distance $H\left(\mathcal{D},\mathcal{D}_y\right)$ by the square root of the variation distance, $\left\|\mathcal{D}-\mathcal{D}_y\right\| = O\left(\varepsilon\sqrt{n}\right)$, we'd only get an upper bound of $O\left(\sqrt{\varepsilon}n^{1/4}\right)$, rather than the $O\left(\varepsilon\sqrt{n}\right)$ that we wanted. To avoid that loss, here we exploited the fact that Lemma 30 upper-bounded the KL-divergence rather than only the variation distance, and we also used Proposition 11, which upper-bounds Hellinger distance directly in terms of KL-divergence, bypassing variation distance.

We now prove the final lemma needed to complete the proof of Theorem 5, by generalizing Lemma 31 from projective measurements to POVMs.

Lemma 32 *If M is any product measurement that is ε-DP on product states, then M is $O\left(\varepsilon\sqrt{n}\right)$-gentle on product states.*

Proof. Again, by Proposition 19, it suffices to restrict attention to pure states. We will give a reduction to the situation already handled in Lemma 31. Suppose we start with the product state

$$|\psi\rangle := |\psi_1\rangle\otimes\cdots\otimes|\psi_n\rangle.$$

Next we apply a POVM to each $|\psi_i\rangle$. This can be modeled as follows: for each i, we apply a unitary transformation U_i to $|\psi_i\rangle$ together with some ancilla qubits that are initially in the state $|0\cdots 0\rangle$. This yields a new state that we can write as

$$|\phi_i\rangle := \sum_{x\in[k]} a_x |x\rangle |v_{ix}\rangle .$$

Here $|x\rangle$ represents a classical computational basis state that the POVM will measure, while $|v_{ix}\rangle$ represents "garbage": some normalized state that depends only on i and x, need not be in the computational basis, and will not be measured.

So now we have

$$|\phi\rangle := |\phi_1\rangle \otimes \cdots \otimes |\phi_n\rangle .$$

Next, we apply our classical algorithm to the basis states $|x_1\rangle, \ldots, |x_n\rangle$, and then we condition on the algorithm outputting y. This yields a new state $|\phi_y\rangle$. What can we say about the relation between $|\phi\rangle$ and $|\phi_y\rangle$?

Let's reorganize $|\phi\rangle$ by collecting $(x_1, \ldots, x_n)$ into a single register that we'll call X, and also collecting all the $|v_{ij}\rangle$'s into a single register that we'll call $|v_X\rangle$. We then have:

$$|\phi\rangle = \sum_{X\in[k]^n} b_X |X\rangle |v_X\rangle ,$$

where $b_X = a_{x_1} \cdots a_{x_n}$, and $|b_X|^2 = \Pr[X]$ is just the probability of X from the perspective of the classical algorithm. By exactly the same reasoning as in the proof of Lemma 31, it follows that

$$|\phi_y\rangle = \sum_{X\in[k]^n:\Pr[X]>0} \frac{b_X}{|b_X|} \sqrt{\Pr[X|y]}\, |X\rangle |v_X\rangle .$$

Therefore,

$$\langle\phi|\phi_y\rangle = \sum_{X\in[k]^n:\Pr[X]>0} b_X^* \frac{b_X}{|b_X|} \sqrt{\Pr[X|y]} = \sum_{X\in[k]^n} \sqrt{\Pr[X]\Pr[X|y]}.$$

So now we have exactly the same expression for the inner product that we had in the proof of Lemma 31. So we can use the same argument to lower-bound the inner product by $1-\varepsilon^2 n$, and to upper-bound both the Hellinger distance and the trace distance between $|\phi\rangle$ and $|\phi_y\rangle$ by $O(\varepsilon\sqrt{n})$.

Finally, recall that $|\phi\rangle$ was obtained by applying a unitary transformation $U = U_1 \otimes \cdots \otimes U_n$ to $|\psi\rangle$ together with some $|0\rangle$ ancilla qubits. Since inner products are unitarily invariant, this means that $U^\dagger |\phi_y\rangle$ also has trace distance at most $O(\varepsilon\sqrt{n})$ from $U^\dagger |\phi\rangle = |\psi\rangle |0\cdots 0\rangle$. Hence we've implemented M in an $O(\varepsilon\sqrt{n})$-gentle manner. ■

Intuitively, what's going on is that the garbage register, $|v_X\rangle$, is completely inert: it's there, but it has no effect on the inner product.

Combining Lemma 28 with Lemma 32 now completes the proof of Theorem 5.

5 Separating Examples

In this section, we prove that the relationships between DP and gentleness notions proved in the preceding two sections are essentially tight, by giving examples of measurements that exhibit their optimality.

5.1 Gentleness to DP

For all $\beta > 0$, let R_β be the "randomized response" algorithm, which for each $i \in [n]$ separately, applies the POVM defined by the matrices

$$E_{\text{reject}} = \begin{pmatrix} \frac{1}{2}+\beta & 0 \\ 0 & \frac{1}{2}-\beta \end{pmatrix}, \quad E_{\text{accept}} = \begin{pmatrix} \frac{1}{2}-\beta & 0 \\ 0 & \frac{1}{2}+\beta \end{pmatrix}$$

to the i^{th} qubit and returns the result. In other words, the output of R_β is an n-bit string, whose i^{th} bit has a bias of β toward the value of the i^{th} qubit in the $\{|0\rangle, |1\rangle\}$ basis. The following is immediate:

Proposition 33 *R_β is ε-DP for $\varepsilon = \ln\left(\frac{1+2\beta}{1-2\beta}\right)$, which is $O(\beta)$ for $\beta \leq \frac{1}{2.01}$, and is not ε'-DP for any $\varepsilon' < \varepsilon$.*

Proof. Flipping the i^{th} input bit can at worst change the probability that the i^{th} output bit assumes some value from $\frac{1}{2}-\beta$ to $\frac{1}{2}+\beta$ (or vice versa), while leaving the other $n-1$ output bits unchanged. ■

We also have:

Proposition 34 *Suppose $n = 1$ (i.e., there is just one qubit). Then R_β is 2β-gentle.*

Proof. Given a qubit in state

$$\rho = \begin{pmatrix} a & b \\ b^* & c \end{pmatrix},$$

here is one way to implement R_β: with probability $1-2\beta$, return 0 or 1 with equal probabilities. With probability 2β, measure ρ in the $\{|0\rangle, |1\rangle\}$ basis and return the result. Suppose without loss of generality that $a > 0$ and we condition on the output being $|0\rangle$. Then the post-measurement state is

$$\sigma := \frac{(1-2\beta)\rho + 2\beta a|0\rangle\langle 0|}{1-2\beta+2\beta a}.$$

The trace distance, $\|\rho-\sigma\|_{\text{tr}}$, can thus be calculated explicitly as

$$\begin{aligned} \frac{2\beta a}{1-2\beta+2\beta a}\|\rho - |0\rangle\langle 0|\|_{\text{tr}} &= \frac{2\beta a\sqrt{|b|^2+c^2}}{1-2\beta+2\beta a} \\ &\leq 2\beta\sqrt{|b|^2+c^2} \\ &\leq 2\beta. \end{aligned}$$

■

Combining Propositions 33 and 34, we get the following corollary:

Corollary 35 *For all $\alpha \in (0,1)$, there exists a measurement that is α-gentle on arbitrary states, but not ε-DP for any $\varepsilon < \ln\left(\frac{1+\alpha}{1-\alpha}\right)$, even on product states.*

Proof. Consider $R_{\alpha/2}$ applied to the first qubit only. ■

This shows that Corollary 24 and Lemma 28 are both tight, up to the factor of 4 in front of the α.

5.2 DP to Gentleness

We now prove that, when we showed that ε-DP on arbitrary states implies $O(\varepsilon n)$-gentleness on arbitrary states (Proposition 27), and that ε-DP on product states implies $O(\varepsilon\sqrt{n})$-gentleness on product states for product measurements (Theorem 5), the n and $\sqrt{n}$ factors were both asymptotically tight.

Recall the measurement L_σ from Sections 1.1 and 1.3, which takes as input an n-qubit state and returns the total Hamming weight, plus a Laplace noise term η of average magnitude σ. We showed, in Proposition 4, that L_σ is $\frac{1}{\sigma}$-DP—and moreover, on *all* n-qubit states, not merely on product states. By contrast, we now observe that L_σ is far from gentle on arbitrary states:

Proposition 36 (Optimality of n Factor) *$L_{n/2}$ is not $\frac{1}{3}$-gentle on n-qubit states.*

Proof. We consider L_σ applied to the mixture

$$\rho := \frac{|0^n\rangle\langle 0^n| + |1^n\rangle\langle 1^n|}{2}.$$

Note that the entire situation is classical, so the question of how L_σ is implemented is irrelevant. Let the measurement outcome be y; then

$$\Pr[y|0^n] = \frac{1}{2\sigma}e^{-|y|/\sigma},$$
$$\Pr[y|1^n] = \frac{1}{2\sigma}e^{-|y-n|/\sigma}.$$

So by Bayes' rule, the post-measurement state is

$$\rho_y = \frac{e^{-|y|/\sigma}|0^n\rangle\langle 0^n| + e^{-|y-n|/\sigma}|1^n\rangle\langle 1^n|}{e^{-|y|/\sigma} + e^{-|y-n|/\sigma}}.$$

Suppose $y \leq 0$. Then we can calculate:

$$\begin{aligned}
\|\rho_y - \rho\|_{\text{tr}} &= \left|\frac{e^{-|y|/\sigma}}{e^{-|y|/\sigma} + e^{-|y-n|/\sigma}} - \frac{1}{2}\right| \\
&= \left|\frac{1}{1+e^{-n/\sigma}} - \frac{1}{2}\right| \\
&= \frac{1}{2}\left(\frac{1-e^{-n/\sigma}}{1+e^{-n/\sigma}}\right) \\
&> \frac{1}{2} - e^{-n/\sigma}.
\end{aligned}$$

If we now make the choice (say) $\sigma = \frac{n}{2}$, we find that this exceeds $\frac{1}{3}$. ∎

It follows that, in going from DP on arbitrary states to gentleness on arbitrary states, we need at least a factor of n blowup in ε; indeed this is true even for product-of-projectives measurements. Hence Proposition 27 is essentially tight.

Likewise, in going from DP on product states to gentleness on product states, we need at least a factor of $\sqrt{n}$ blowup in ε, and this is true even for product-of-projectives measurements. Hence Lemma 31 is essentially tight. The example that shows this is again L_σ, albeit this time with $\sigma = \sqrt{n}$:

Proposition 37 (Optimality of $\sqrt{n}$ Factor) *$L_{\sqrt{n}}$ is not α-gentle on n-qubit product states, for any $\alpha = o(1)$.*

Proof. Let $\sigma = \sqrt{n}$, and consider L_σ applied to the uniform distribution $\mathbb{I}/2^n$. Again, since the entire situation is classical, the question of how L_σ is implemented is irrelevant. Let the measurement outcome be y; then by Bayes' rule, the post-measurement state is

$$\rho_y = \frac{\sum_{X \in \{0,1\}^n} e^{-|y-|X||/\sigma} |X\rangle\langle X|}{\sum_{X \in \{0,1\}^n} e^{-|y-|X||/\sigma}}.$$

So suppose $y \leq 0$, and assume without loss of generality that n is odd. Then we can calculate:

$$\begin{aligned}
\left\|\rho_y - \rho\right\|_{\text{tr}} &= \frac{1}{2} \sum_{X \in \{0,1\}^n} \left| \frac{e^{-|y-|X||/\sigma}}{\sum_{Z \in \{0,1\}^n} e^{-|y-|Z||/\sigma}} - \frac{1}{2^n} \right| \\
&= \frac{1}{2} \sum_{k=0}^{n} \binom{n}{k} \left| \frac{e^{-|y-k|/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-|y-\ell|/\sigma}} - \frac{1}{2^n} \right| \\
&= \frac{1}{2} \sum_{k=0}^{n} \binom{n}{k} \left| \frac{e^{-k/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} - \frac{1}{2^n} \right| \\
&\geq \frac{1}{2} \left| \sum_{k=0}^{(n-1)/2} \binom{n}{k} \frac{e^{-k/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} - \sum_{k=(n+1)/2}^{n} \binom{n}{k} \frac{e^{-k/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} \right| \\
&= \frac{1}{2} \left| \sum_{k=0}^{(n-1)/2} \binom{n}{k} \frac{e^{-k/\sigma} - e^{-(n-k)/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} \right| \\
&\geq \frac{1}{2} \left| \sum_{k=0}^{n/2-\sqrt{n}} \binom{n}{k} \frac{e^{-k/\sigma} - e^{-\left(n/2+\sqrt{n}\right)/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} \right| \\
&\geq \frac{1}{4} \left| \sum_{k=0}^{n/2-\sqrt{n}} \binom{n}{k} \frac{e^{-k/\sigma}}{\sum_{\ell=0}^{n} \binom{n}{\ell} e^{-\ell/\sigma}} \right| \\
&\geq \frac{1}{4} \left| \sum_{k=0}^{n/2-\sqrt{n}} \binom{n}{k} \frac{1}{2^n} \right| \\
&= \Omega(1).
\end{aligned}$$

■

6 Shadow Tomography

Having developed the connection between DP and gentleness, we're now ready to apply the connection to shadow tomography. First, in Section 6.1, we review a recent algorithm of Aaronson et al. [7] for online learning of quantum states, which we'll need as a central ingredient. Then, in Section 6.2, we present and analyze our new *Quantum Private Multiplicative Weights (QPMW)* algorithm,

which builds on the Private Multiplicative Weights (PMW) algorithm of Hardt and Rothblum [27]. QPMW proves Theorem 9: that is, it shows that it's possible to do shadow tomography using only $O\left((\log m)^2 (\log d)^2 / \varepsilon^8\right)$ copies of an unknown mixed state ρ, where m is the number of known accept/reject measurements, d is the dimension of ρ, and ε is the accuracy with which we want to estimate each measurement's acceptance probability—in a way that, moreover, is *online* (i.e., processes the measurements one at a time) and ε-gentle (i.e., damages the copies of ρ by at most ε in trace distance).

6.1 Online Learning of Quantum States

Aaronson et al. [7] recently defined and studied the problem of *online learning of quantum states*. Here we have an unknown d-dimensional mixed state ρ, and a learner is presented with a sequence $E_1, E_2, \ldots$ of two-outcome POVM measurements. For each measurement E_t, the learner tries to anticipate $\mathrm{Tr}(E_t \rho)$, the probability that E_t accepts ρ, up to accuracy $\pm\varepsilon$. Indeed, the learner maintains a "hypothesis state" σ_t, and on each measurement E_t, if the hypothesis differs appreciably from the unknown state ρ with respect to this measurement—that is, if

$$|\mathrm{Tr}(E_t \rho) - \mathrm{Tr}(E_t \sigma_{t-1})| > \varepsilon$$

—then we say that the learner was "wrong," and we allow it to *update* its state by giving it an approximation $b_t \in [0,1]$ to the correct answer, where (say) $|\mathrm{Tr}(E_t \rho) - b_t| \leq \frac{\varepsilon}{10}$. The learner's goal is to upper-bound the total number of times that it's ever wrong, even assuming that the sequence of E_t's and b_t's is chosen adaptively, by an adversary who sees the learner's hypotheses.

Perhaps surprisingly, Aaronson et al. [7] showed that the total number of mistakes can be upper-bounded by $O\left(\frac{\log d}{\varepsilon^2}\right)$—so for example, only $O\left(\frac{n}{\varepsilon^2}\right)$ for a state of n qubits (even though the state space has dimension 2^n).

We observe that the same bound holds even under a slight relaxation of the update condition: namely, updates can also be triggered when the hypothesis has error between $\frac{\varepsilon}{3}$ and ε. If an update is triggered, then the learner again receives an $\frac{\varepsilon}{10}$-approximation to the correct answer.

Theorem 38 (Variant: Online Learning of Quantum States [7, Theorem 1]) *There is an explicit procedure for online learning of quantum states that makes at most $\ell(d, \varepsilon) = O\left(\frac{\log d}{\varepsilon^2}\right)$ updates, so long as updates never occur when the hypothesis has error smaller than $\frac{\varepsilon}{3}$, and updates always occur when the error is ε or larger.*

We emphasize that when the error is in the range $[\frac{\varepsilon}{3}, \varepsilon)$, updates may or may not occur.

Aaronson et al. [7] actually gave two explicit procedures that achieve the above bound: one based on online convex optimization, the other on matrix multiplicative weights. Both procedures use an amount of computation per measurement that's polynomial in d.

In this work, however, we'll be able simply to use Theorem 38 as a black box. We'll view an online learning procedure as specified by its initialization procedure, which outputs an initial hypothesis state $\sigma_0 \leftarrow \mathbf{OnlineLearn}(d)$, and an update procedure used to update the hypothesis state $\sigma_t \leftarrow \mathbf{OnlineUpdate}(\sigma_{t-1}, b_t)$.

6.2 Online Shadow Tomography

Our Quantum Private Multiplicative Weights (QPMW) algorithm for gentle online shadow tomography is presented in Figure 1.

Parameters: Intended number of queries $m \in \mathbb{N}$, gentleness and accuracy parameters $\alpha, \varepsilon, \delta > 0$ and noise magnitude $\mu > 0$ (set in the proof below, see Equation 10).
Input: n and a product state $\rho = \rho_1 \otimes \cdots \otimes \rho_n$, where the ρ_i's are d-dimensional mixed states.

Algorithm:
Initialize the online learner $\sigma_0 \leftarrow \textbf{OnlineLearn}(d)$
In each round $t \leftarrow 1, 2 \ldots, m$, when receiving two-outcome measurement E_t, do the following:

1. Apply the two-outcome **CheckForUpdate** measurement:
   (a) Apply the Laplace measurement with noise magnitude $n\mu$ to the n registers to compute (but *do not measure*) a noisy estimate a_t of $\mathrm{Tr}(E_t\rho)$
   (b) Compute *and measure* the decision bit u_t, which is 1 if and only if $|a_t - \mathrm{Tr}(E_t\sigma_{t-1})| > \frac{\varepsilon}{2}$
   (c) Uncompute the noise and intermediate values computed in the above measurement
2. If $u_t = 0$ (no update), then set $\sigma_t \leftarrow \sigma_{t-1}$ and output the answer $b_t \leftarrow \mathrm{Tr}(E_t\sigma_{t-1})$
3. Otherwise ($u_t = 1$, i.e. an "update round"):
   (a) If there have already been $\ell(d, \varepsilon)$ prior update rounds (see Theorem 38), then abort
   (b) Apply the Laplace measurement with noise magnitude $n\mu$ to the n registers to compute a noisy estimate b_t of $\mathrm{Tr}(E_t\rho)$, uncompute the noise and intermediate values computed within this measurement
   (c) Run a round of online learning: $\sigma_t \leftarrow \textbf{OnlineUpdate}(\sigma_{t-1}, b_t)$
   (d) output the answer b_t

Figure 1: QPMW Algorithm

Theorem 39 *Let $\alpha, \beta, \varepsilon, \delta > 0$ be gentleness and accuracy parameters. There exists a setting for the noise magnitude μ for which the online shadow tomography algorithm presented in Figure 1 is (α, δ)-gentle. Moreover, given sufficiently many copies n, where*

$$n = O\left(\frac{\left(\log^2 m + \log\frac{1}{\delta}\right) \cdot \log^2 d \cdot \log\frac{1}{\beta}}{\varepsilon^6 \min\{\alpha, \varepsilon\}^2}\right),$$

the algorithm's error is bounded by ε with probability at least $1 - \beta$ over its coins and its measurements.

Proof. We first prove gentleness and then turn our attention to bounding the error (the accuracy proof builds on the algorithm's gentleness).

Gentleness. Note that we argue gentleness for *any* product state provided as input (i.e., for gentleness, we don't assume that the input is n copies of a single state). By Proposition 19, it suffices to consider the case where the input is a pure product state $|\psi\rangle = |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle$. It is straightforward to see that the update rounds are gentle: we run two DP measurements in each update round, and their outcomes are gentle by Theorem 5. This is stated below in Claim 40. The non-update rounds are certainly no less gentle than the update rounds (after all, we only run the first measurement), but we expect to have a very large number of no-update rounds, and so we need a much better bound. We obtain such a bound by restricting our attention to the damage that can be caused by the conditioned superoperator **CheckForUpdate**, conditioned on the output being 0 (no update). One important challenge is showing that the damage (conditioned on this particular outcome) is tightly related to the probability of an update. Thus, it will be highly unlikely for a sequence of rounds (even a very long sequence!) to cause significant damage before it triggers an update. The second challenge is bounding the damage that can be caused by a sequence of *conditioned* superoperators. This is done via a delicate accounting argument, which relies on Lemma 17

We begin by fixing some notation. First, given a superoperator and a fixed output y, we use the term *conditioned superoperator* to refer to running the superoperator conditioned on the output being y.[17] The QPMW algorithm's output in any run can be specified by $m' \leq m$, the number of rounds before an "abort" (if any) occurs, and by a vector $\bar{y}$ of outcomes, where for each $t \in \{1, \ldots, m'\}$, the outcome in round t is $y_t \in [0,1] \cup \{\perp\}$. In no-update rounds the outcome is $\perp$, while in update rounds the outcome $y_t = b_t$ is the noisy answer returned by the algorithm. Note that m' and the vector $\bar{y}$ of outcomes indeed specify all outputs of the algorithm. For an intermediate round $t \in [1, m']$, we can also consider the vector $\bar{y}_{\leq t}$ of outcomes in the first t rounds. Taking $|\psi\rangle$ to be the initial state of the algorithm, we take $|\psi_{\leq t}\rangle$ to be the state after round t, conditioned on the outcomes $\bar{y}_{\leq t}$ (and given the measurements $E_1, \ldots, E_t$). The initial state is thus $|\psi\rangle = |\psi_{\leq 0}\rangle$, and the final state is $|\psi_{\leq m'}\rangle$.

Consider an execution of the algorithm at the beginning of the t^{th} round. The outcomes in previous rounds are given by $\bar{y}_{\leq t-1}$, which determines the learned state σ_{t-1}. Let E_t be the t^{th} measurement. We define λ_t to be the probability that the **CheckForUpdate** $(E_t, |\psi\rangle, \sigma_{t-1})$ measurement returns 1, i.e. the probability of an update on the *original* state $|\psi\rangle$. Similarly, we take κ_t to be the probability that **CheckForUpdate**$(E_t, |\psi_{\leq t-1}\rangle, \sigma_{t-1})$ returns 1, i.e. the probability of an update on the real state in the registers at the beginning of the t^{th} round.

The following claims bound the damage that can occur if we run the t^{th} round with a fresh copy of the original state in the registers.

Claim 40 *Every round of the algorithm is an $O\left(\frac{1}{\sqrt{n\mu}}\right)$-gentle superoperator.*

Claim 41 *Take n and μ to be set as in Equations 11 and 10. Let $|\psi_{no}\rangle$ be the state after we run* **CheckForUpdate**$(E_t, |\psi\rangle, \sigma_{t-1})$*, and condition on the output 0 ("no update," which occurs with probability $1 - \lambda_t$). We have:*

$$\left\| |\psi\rangle\langle\psi| - |\psi_{no}\rangle\langle\psi_{no}| \right\|_{\mathrm{tr}} = O\left(\frac{\lambda_t}{\sqrt{n\mu}}\right).$$

[17] In the terminology of Section 2.3, a conditioned superoperator is a quantum operation but where we normalize the output state.

Claim 40 follows immediately from the differential privacy of the Laplace measurement and from Theorem 5. We defer the proof of Claim 41, which is technically involved and lengthy (see below). We first show that, given this claim, the algorithm (taken as a whole) is gentle.

Epoch superoperators. For the analysis, we divide an execution of the algorithm into epochs, where each epoch is comprised of one or more rounds. The k^{th} epoch begins in round t_k (where $t_1 = 1$). The k^{th} epoch ends on the first round $t' \geq t_k$ where one of the following occurs:

(1) An update happens (or the epoch reaches the last round m).

(2) The probability of an update, if each round was run on the original state, becomes too large:

$$\prod_{j=t_k}^{t'} (1 - \lambda_j) \leq \frac{1}{2}. \tag{6}$$

Naturally, the last epoch always ends on the last round m'. The crux of the gentleness analysis is bounding the damage done to the state within any single epoch. A separate argument shows that the number of epochs cannot be too large.

Viewing each epoch as a superoperator, it is specified by a list of measurements $E_{t_k}, E_{t_k+1}, \ldots$ that would be chosen so long as no updates occurred. Note that this list is indeed fixed: while the strategy that chooses the actual measurements $E_{t_k}, E_{t_k+1}, \ldots$ can be adaptive, it specifies a fixed sequence of measurements (known in advance) that will be chosen so long as the outputs are "no update." Let $t'' \geq t_k$ be the first round that meets Condition (6). The epoch processes the list of measurements $E_{t_k}, E_{t_k+1}, \ldots, E_{t''}$ until an update occurs (or the last measurement in this list is processed). Note that t'' depends on the initial state $|\psi\rangle$, but it is fixed in advance. Given the list of measurements $E_{t_k}, E_{t_k+1}, \ldots$, the output of the epoch superoperator is the length a list of "no update" decisions of length $s \in [0, t'' - t_k]$, followed (if an update occurs in the final round) by the output b_{t_k+s} of the Laplace measurement used to approximate the value of E_{t_k+s}.

We bound the damage that can be caused to the original (product) state $|\psi\rangle$ by running the epoch superoperator. We also show that running the epoch superoperator on $|\psi\rangle$ triggers an update with constant probability, but with constant probability no update occurs before round t''.

Claim 42 *There exists a noise magnitude $\mu = O\left(\frac{1}{\sqrt{n}\varepsilon^2}\right)$ such that the following holds. Fixing any round $t_k \in [m]$, prior measurements $E_1, \ldots, E_{t_k-1}$, and a history of outputs $\bar{y}_{\leq t_k-1}$ in the previous rounds, define the epoch superoperator as above. Then:*

(1) When we run the epoch superoperator on the state $|\psi\rangle$, the probability that an update occurs is at least 0.15*.*

(2) When we run the epoch superoperator on the state $|\psi\rangle$, the probability that no *update occurs before the round t'' is at least* 0.4*.*

(3) Let $|\psi'\rangle$ be the state in the registers after running this superoperator on the original state $|\psi\rangle$ (including observing the epoch's outputs). The damage is bounded by:

$$\left\| |\psi\rangle\langle\psi| - |\psi'\rangle\langle\psi'| \right\|_{\mathrm{tr}} = O\left(\frac{1}{\sqrt{n}\mu}\right).$$

Proof. For any possible last round $t' \geq t_k$, and any possible output y of the epoch superoperator (comprised of a sequence of no-updates, which may or may not end with an update), we bound the damage as follows. We take $t'' \geq t'$ to be the round on which the epoch always ends (unless there is an earlier update). Since Condition (6) did not hold at the beginning of round $t' - 1$, we have:

$$p = \prod_{j=t_k}^{t'-1} (1 - \lambda_j) \geq \frac{1}{2}.$$

Using the fact that for any $\xi \in [0, 1]$, we have that $0 \leq 1 - \xi \leq e^{-\xi}$:

$$\frac{1}{2} \leq p = \prod_{j=t_k}^{t'-1} (1 - \lambda_j) \leq \prod_{j=t_k}^{t'-1} e^{-\lambda_j} = e^{-\sum_{j=t_k}^{t'-1} \lambda_j}.$$

By taking logarithms on both sides of the above inequality we get:

$$\sum_{j=t_k}^{t'-1} \lambda_j < 1. \tag{7}$$

Claim 41 gives a bound ε_j on the damage when running the j^{th} **CheckForUpdate** conditioned superoperator (on the original state), conditioned on output 0. Recall that this bound ε_j is linear in the update probability λ_j. Claim 40 gives a bound ε_{t_k+s} on the damage caused by the conditioned superoperator run in the last round, conditioned on any possible outcome in that round. Combining these bounds with inequality (7), we get:

$$\sum_{j=t_k}^{t'} \varepsilon_j = O\left(\frac{1}{\sqrt{n}\mu}\left(1 + \sum_{j=t_k}^{t'-1} \lambda_j\right)\right) = O\left(\frac{1}{\sqrt{n}\mu}\right).$$

Define q to be the probability of no update in rounds $t_k, \ldots, t' - 1$ in a "real" execution of the epoch superoperator on the state $|\psi\rangle$ (and note that $q > 0$, because we are considering an output y that can actually occur). Applying Lemma 17 to the conditioned superoperator's run in the first $t' - t_k - 1$ rounds, and using also the bound on μ in the claim's statement, we get:

$$|p - q| \leq \sum_{j=t_k}^{t'-1} \varepsilon_j < 0.1, \tag{8}$$

which in particular implies that $q \geq 0.4$, proving item (2) above. By Lemma 17 (see also the remark following that lemma about composing with a final superoperator—in our case, the $(t')^{th}$ round), we conclude that:

$$\left\| |\psi\rangle\langle\psi| - |\psi'\rangle\langle\psi'| \right\|_{\mathrm{tr}} \leq 5 \sum_{j=t_k}^{t'} \varepsilon_j = O\left(\frac{1}{\sqrt{n}\mu}\right).$$

Bounding the update probability. To lower-bound the probability of an update, observe first that if the probability $\lambda_{t''}$ of an update in the last round, when we run it on a fresh copy of the

state $|\psi\rangle$, satisfies $\lambda_{t''} \geq \frac{1}{4}$, then by gentleness of the epoch superoperator as a whole (see above), when we run it on $|\psi\rangle$, the probability of an update in the last round (run on the state $|\psi_{\leq t''}\rangle$) is greater than 0.15.

Thus, we restrict our attention to the case that $\lambda_{t''} < \frac{1}{4}$. Since t'' is the first round where Condition (6) is violated, we know that $\prod_{j=t_k}^{t''} (1-\lambda_j) \leq \frac{1}{2}$. I.e., we have a lower bound on the probability of an update if each round was run on a fresh copy of $|\psi\rangle$. Since we assume $\lambda_{t''} < \frac{1}{4}$, we in fact have an upper bound on the probability of no update in the first $t''-1$ rounds of such an execution:

$$p = \prod_{j=t_k}^{t''-1} (1-\lambda_j) \leq \frac{3}{4}.$$

By equation (8) (restricted to the case $t' = t''$), we deduce a similar bound on the probability q of no update in the first $t''-1$ rounds of the actual execution (an execution that does not get fresh copies of $|\psi\rangle$). In particular, the probability of no update in this "actual" execution is at most 0.85. ■

Accumulated damage. By Claim 42, running each epoch superoperator on the initial state only results in bounded damage, and triggers an update with constant probability. By Lemma 15 (additivity of damage), when we run a sequence of k epochs, the total damage is at worst multiplied by k. Moreover, so long as this accumulated damage is smaller than 0.05, each epoch still triggers an update with probability at least 0.1 (because the trace distance between the original state and the state in the registers when we run the epoch is bounded). Under these conditions, by Azuma's inequality, with all but $\frac{\delta}{2}$ probability, the number of epochs that occur before $\ell(d,\varepsilon)$ updates are triggered (and the QPMW algorithm aborts) is at most:

$$k = O\left(\ell(d,\varepsilon) + \sqrt{\ell(d,\varepsilon)\log\frac{1}{\delta}}\right).$$

By Theorem 38 we have that $\ell(d,\varepsilon) = O\left(\frac{\log d}{\varepsilon^2}\right)$. Note that the choice of noise parameter μ guarantees that the accumulated damage over k such rounds is indeed less than 0.05 (in fact it is less than α; see equation (9)). We conclude that in this random process, the probability that each epoch triggers an update stays above 0.1 for the first k epochs.

By Claim 42 and Lemma 15 (additivity of damage), we can bound the total damage by the number of epochs times the damage per epoch, and we get that with all but $\frac{\delta}{2}$ probability over the coins and measurements made by QPMW:

$$\begin{aligned}\left\| |\psi\rangle\langle\psi| - |\psi_{\leq m'}\rangle\langle\psi_{\leq m'}| \right\|_{\mathrm{tr}} &= O\left(\sum_{j=1}^{k} \frac{1}{\sqrt{n}\mu}\right) \\ &= O\left(\frac{\log d}{\sqrt{n}\mu\varepsilon^2} + \frac{\sqrt{\log d \log\frac{1}{\delta}}}{\sqrt{n}\mu\varepsilon}\right).\end{aligned}$$

Accuracy. For given gentleness and accuracy parameters $\alpha, \delta, \varepsilon > 0$, we fix the noise parameter μ and then analyze the number of copies needed to guarantee accuracy with high probability. We

assume without loss of generality that $\alpha \leq \varepsilon/100$ (if a larger α is specified, we simply run the algorithm with $\alpha' = \varepsilon/100$). We set the parameters so that in an "ideal" run of the algorithm, where each round is run on a fresh copy of the state $\rho^{\otimes n}$, the algorithm is $\varepsilon/10$-accurate with all but a small constant probability. We then use the algorithm's gentleness to show that this implies accuracy in "real" runs of the algorithm: namely, we show that in a real run, the algorithm is ε-accurate with all but a small constant probability. The error probability can be reduced by independent repetitions.

We begin by setting the parameters so that with high probability, the total damage to the state is bounded by α, and recall also that we assume $\alpha \leq \frac{\varepsilon}{100}$. This imposes a constraint on μ:

$$\min(\alpha, \varepsilon) = O\left(\frac{\log d}{\sqrt{n}\mu\varepsilon^2} + \frac{\sqrt{\log d \log \frac{1}{\delta}}}{\sqrt{n}\mu\varepsilon}\right),$$

or equivalently:

$$n = \Omega\left(\left(\frac{1}{\mu\varepsilon \min(\alpha, \varepsilon)}\left(\frac{\log d}{\varepsilon} + \sqrt{\log d \log \frac{1}{\delta}}\right)\right)^2\right). \tag{9}$$

Note that this setting also satisfies the conditions of Claim 42.

We also want to guarantee that with high probability, an ideal run of the algorithm would give accurate answers. This imposes an upper bound on the noise magnitude μ. We analyze the accuracy by dividing the execution into epochs, as was done in the gentleness analysis above.

Claim 43 (Ideal run accuracy) *Consider an ideal run of the algorithm (where each round is run on a fresh copy of $\rho^{\otimes n}$) where we set:*

$$\mu = O\left(\frac{\varepsilon}{\log m}\right). \tag{10}$$

Consider an epoch that can run for at most s rounds. The following all hold:

(1) With all but $\frac{s}{1000m}$ probability, there will not *be an update in any round t of the epoch where $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \leq \frac{\varepsilon}{3}$.*

(2) If in any round t of the epoch it is the case that $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \geq \varepsilon$, then an update occurs in that round with all but $\frac{1}{1000m}$ probability (note this condition can only hold on the round that always ends the epoch).

(3) If the epoch ends in an update round, then the noisy answer b_t is $\left(\frac{\varepsilon}{10}\right)$-accurate with all but $\frac{1}{1000m}$ probability.

■

Proof. The claim follows immediately from the exponential tails of the Laplace distribution: in each round, for each draw of Laplace noise, with all but $\frac{1}{1000m}$ probability, the noise magnitude is at most $\frac{\varepsilon n}{10}$.

Recall that an epoch can end before reaching its last round. However, the probability of each epoch reaching its final round is at least 1/2 (by the definition of the epoch superoperator). Thus, if an epoch can run for at most s rounds, then the *expected* number of rounds is at least $s/2$.

We conclude that with probability at least 0.9, the sum, over all epochs, of the number of rounds for which each epoch can run, is at most $10m$ (by Markov's inequality). By Claim 43, taking a union bound over all epochs, and taking μ as set as in Equation 10, we deduce that with all but a small constant probability over the noise choices, the conditions of the online learning theorem for quantum states (Theorem 38) all hold in all rounds simultaneously. By that theorem, we conclude that with all but small constant probability over its coins, the QPMW algorithm does not abort, and its answers are all ε-accurate.

How many copies do we need? Before proceeding to prove that a real run of the algorithm is also accurate, we specify the number of copies needed to simultaneously satisfy the constraints in equations (9) and (10) by taking n to be large enough. We can do so while still guaranteeing the upper bound:

$$n = O\left(\frac{\log^2 m}{\varepsilon^4 \min(\alpha,\varepsilon)^2} \cdot \left(\frac{\log^2 d}{\varepsilon^2} + (\log d)\left(\log\frac{1}{\delta}\right)\right)\right). \tag{11}$$

Note that this setting of n, which we use in the proof of Claim 41, also guarantees that $\sqrt{n}\mu$ is a sufficiently large constant. Further, this number of copies guarantees accuracy with all but small constant probability. The error probability can be reduced to β by running $O(\log(1/\beta))$ independent copies of the algorithm, and outputting the median answer in each round.

For simplicity, in the statement of Theorem 39 we claim a slightly more relaxed bound of:

$$n = O\left(\frac{\left(\log^2 m + \log\frac{1}{\delta}\right)\cdot \log^2 d \cdot \log\frac{1}{\beta}}{\varepsilon^6 \min(\alpha,\varepsilon)^2}\right).$$

A hybrid execution. Consider a hybrid execution, where each epoch superoperator (see above) is run on the "real" state (with no substitutions), but after each superoperator completes its operation, we replace the resulting state with a fresh copy of $\rho^{\otimes n}$ before proceeding to the next epoch superoperator. Since each epoch is $\alpha_e = O\left(\frac{1}{\sqrt{n}\mu}\right)$-gentle (Claim 42), we can apply the Damage Lemma (Lemma 17) to conclude accuracy properties for the epoch:

Claim 44 (Hybrid run accuracy) *Consider a hybrid run of the algorithm (where each epoch is run on a fresh copy of $\rho^{\otimes n}$), with the parameters set as in Equations* (9), (10), *and* (11). *Let α_e be the bound on the gentleness of each epoch. Consider an epoch that can run for at most s rounds. The following all hold:*

(1) With all but $\frac{s}{1000m} + \alpha_e$ probability, there will not *be an update in any round t of the epoch where $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \leq \frac{\varepsilon}{3}$.*

(2) If in the final round of the epoch it is the case that $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \geq \varepsilon$, then an update occurs in that round with all but $\frac{1}{1000m} + \alpha_e$ probability.

(3) If the epoch ends in an update round, then the noisy answer b_t is $\frac{\varepsilon}{10}$-accurate with all but $\frac{1}{1000m} + \alpha_e$ probability.

Proof. Consider the set I of rounds where $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \leq \frac{\varepsilon}{3}$. By Claim 43, the probability that in an ideal execution an update occurs in one of the rounds in I is at most $\frac{s}{1000m}$. Applying Lemma 17 to the epoch superoperator, we conclude that the probability an update occurring in one of the rounds in I is at most $\frac{s}{1000m} + \alpha_e$. We note that in this application of Lemma 17, we restrict to the subset of quantum operations corresponding to rounds in I (and condition on the "no update" outcome in those rounds). Claim 43 further bounds the ideal-execution probability of no update if in the last round $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \geq \varepsilon$, and the probability that the update ends in an update round but the noisy answer is not $\frac{\varepsilon}{10}$-accurate. By α_e-gentleness of the epoch superoperator, we conclude that the probabilities of these two events occurring in the hybrid execution are both bounded by $\frac{1}{1000m} + \alpha_e$. ■

By Claim 42, the probability that there is no update until the last (s^{th}) round of an epoch is at least 0.4. Thus, in the hybrid execution, the *expected* number of rounds for which an s-round epoch will run is at least $0.4s$. Similarly to the analysis of the ideal execution, taking a union bound over k epoch superoperators and taking $\tau > 0$ to be a small constant, we conclude that with all but $\tau + O(\alpha_e \cdot k)$ probability, the conditions of the online learning theorem all hold and the answers returned are all ε-accurate. Further, by the choice of parameters in Equation (9), we know that with high probability, when we run QPMW and take k^* to be the number of epochs needed to process all m measurements, we have $O(\alpha_e \cdot k^*) = O(\alpha)$. We conclude that with all but a small constant probability, a hybrid execution of QPMW does not terminate prematurely, and is ε-accurate on every measurement.

The real execution. Lastly, we consider the real execution, where the epoch superoperators are run in sequence, without any refreshing of the state in the registers. We use the gentleness of the epoch superoperator to conclude that the algorithm remains accurate in its real execution.

Claim 45 (Real run accuracy) *Consider a real run of the algorithm, with the parameters set as in Equations* (9)*,* (10)*, and* (11)*. With all but small constant probability over the algorithm's coins, the following hold in every round t of the algorithm (simultaneously):*

(1) If $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \leq \frac{\varepsilon}{3}$, then there is no update.

(2) If $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \geq \varepsilon$, then there is an update.

(3) If t is an update round, then the noisy answer b_t is $\frac{\varepsilon}{10}$-accurate.

Proof. Let B be the ("bad") event that in some round t of QPMW it is either the case that:

(i) an update occurs even though $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \leq \frac{\varepsilon}{3}$, or

(ii) no update occurs even though $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\sigma_{t-1})| \geq \varepsilon$, or

(iii) t is an update round, and the noisy answer b_t is *not* $\frac{\varepsilon}{10}$-accurate.

By the foregoing analysis, the probability of the event B in the hybrid execution is bounded by a small constant, say τ. We would like to now make a similar argument for a real execution, where the state is not "refreshed" between epoch superoperators.

Towards this, let k be a bound on the number of epoch superoperators in a run of QPMW, and let α_e be the bound on the gentleness of each epoch superoperator. We consider further hybrids,

where in the i^{th} hybrid $\mathcal{H}_i$, the first i epochs are each run on fresh copies of $\rho^{\otimes n}$, but there is no further refreshing after the i^{th} epoch. Thus the first hybrid $\mathcal{H}_1$ equals the real execution, and the k^{th} hybrid $\mathcal{H}_k$ equals the hybrid execution. By α_e-gentleness of the epoch superoperator, we have that for every i:

$$\left| \Pr_{\mathcal{H}_i}[B] - \Pr_{\mathcal{H}_{i+1}}[B] \right| \leq \alpha_e.$$

This is simply because the i^{th} and $(i+1)^{st}$ hybrid differ only in running the $(i+1)^{st}$ epoch: in $\mathcal{H}_i$ that epoch is run on the state in the registers after the i^{th} epoch, whereas in $\mathcal{H}_{i+1}$ that epoch is run on a fresh copy of $\rho^{\otimes n}$. By the α_e-gentleness of the i^{th} epoch, the trace distance between these two states is at most α_e. So the two hybrids only differ in the probability that the event B occurs in the $(i+1)^{st}$ epoch, and this difference in probabilities is upper-bounded by α_e.

By a hybrid argument, we conclude that the probability of the event B occurring in the real execution is at most $\tau + k \cdot \alpha_e$. Further, by the choice of parameters in Equation (9), we know that we can take k^* to be a bound on the number of epochs such that with high probability, k^* epochs suffice to process all m measurements, and $O(\alpha_e \cdot k^*) = O(\alpha)$. We conclude that with all but a small constant probability, the real execution of QPMW does not terminate prematurely, and is ε-accurate on every measurement. ■

Finally, we reduce the error probability to β by running $O(\log(1/\beta))$ independent executions and outputting the median answer in each round. This completes the accuracy proof for QPMW.

■

Proof of Claim 41. We begin by assuming that the probability λ_t of an update is smaller than some sufficiently small constant. If this is not the case, then the claim follows immediately from Lemma 32, because **CheckForUpdate** runs a $\frac{1}{n\mu}$-DP classical algorithm. Further, we assume throughout that $\sqrt{n}\mu$ is larger than a sufficiently large constant (see the remark following equation (11)).

We follow similar reasoning to the proof of Lemma 32. We begin with a pure product state in the registers

$$|\psi\rangle := |\psi_1\rangle \otimes \cdots \otimes |\psi_n\rangle .$$

Let $|\psi_{\text{no}}\rangle$ be the state after applying the conditioned superoperator **CheckForUpdate**, conditioned on $u_t = 0$ ("no update"). The **CheckForUpdate** superoperator applies the POVM E_t to each $|\psi_i\rangle$, and then runs a classical DP algorithm on the n bits observed. To implement it, we first apply a unitary transformation (to the state and ancilla qubits). This gives a new state:

$$|\phi_i\rangle := \sum_{b_i \in \{0,1\}} a_{i,x} |b_i\rangle |v_{i,b_i}\rangle .$$

Let $X \in \{0,1\}^n$ be the values observed when measuring the registers $|b_i\rangle$. We draw a noise value η from the Laplace distribution with magnitude $n\mu$, and output $u_t = 0$ (no update) whenever:

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i + \frac{\eta}{n} - \operatorname{Tr}\left(E_t \omega_{t-1}\right) \right| \leq \frac{\varepsilon}{2}.$$

Let $\mathcal{D}$ be the distribution over $X \in \{0,1\}^n$ defined by $\Pr_{\mathcal{D}}[X] = |\alpha_X|^2$, where $\alpha_X = a_{i,1} \cdots a_{i,n}$, and note that $\mathcal{D}$ is indeed a product distribution. Let $\mathcal{D}_0$ be the distribution $\mathcal{D}$ conditioned on the

event when we run the above (classical) procedure on X we get $u_t = 0$ (no update). Following the proofs of Lemmas 32 and 31, we can implement the **CheckForUpdate** measurement so that:

$$\| |\psi\rangle\langle\psi| - |\psi_{\text{no}}\rangle\langle\psi_{\text{no}}| \|_{\text{tr}} \leq \sqrt{2 \operatorname{KL}(\mathcal{D}, \mathcal{D}_0)}. \tag{12}$$

At this point we diverge from the proof of Lemma 32. There, we considered the distribution $\mathcal{D}_y$ obtained by conditioning the product distribution $\mathcal{D}$ on an outcome y of a σ-DP algorithm. We bounded the KL-divergence between these distributions, and used that to bound the trace distance by $O(\sigma\sqrt{n})$. Here, while we know that the **CheckForUpdate** measurement is σ-DP for $\sigma = \frac{1}{n\mu}$, when the probability λ_t of an update is much smaller than $\frac{1}{\sqrt{n}\mu}$, we want to argue that observing a "no update" answer causes much less damage to the state.

Improving the DP guarantee. The intuition is that when λ_t is small, for a "typical" input X drawn from $\mathcal{D}$, the probability of no update is quite large: $1 - \lambda_t$. For an adjacent input X', this probability of no update is at least $1 - e^{\sigma}\lambda_t$. For small λ_t, the log-ratio between these two probabilities is roughly $\lambda_t\sigma$. A compelling strategy is to try to bound the KL-divergence using this improved bound, by following a similar argument to the proof of Lemma 32. For observe that that proof applies even when we focus on any particular output y—in this case, "no update"—using the log-ratio guaranteed for that particular output.

The catch, which significantly complicates the proof, is that not all inputs drawn from $\mathcal{D}$ are "typical." Some of these inputs have much higher update probabilities than λ_t, whereas the proof of Lemma 32 required a worst-case bound that applies to *every* input in the support of $\mathcal{D}$. On the other hand, by concentration bounds on the Hamming weights of inputs drawn from $\mathcal{D}$, the probability of drawing an X for which the update probability is significantly higher than λ_t is very small.

To obtain an improved bound, we extend the proof of Lemma 32 to this case, using concentration of the (generalized) binomial distribution (a subgaussian distribution), to show that while the contribution of "far" inputs to the KL-divergence grows, their probability shrinks more quickly than this growth. To do this, we partition the inputs into disjoint sets Δ_j, according to the difference between their Hamming weight and the expected Hamming weight. We account for the contributions of each set in this partition to the KL-divergence to show the claimed bound. The details (which can get long and technical) follow.

The event Δ_j. For each integer $j \in [1, \sqrt{n}]$, we define the event $\Delta_j \subseteq \{0,1\}^n$ to consist of all inputs whose Hamming weights are at least $(j-1)\sqrt{n}$ and less than $j\sqrt{n}$ away from the expectation:

$$\Delta_j = \left\{ X \in \{0,1\}^m : \left| \left(\sum_{i=1}^{n} X_i \right) - \mathrm{E}_{Y \sim \mathcal{D}} \left[\sum_i Y_i \right] \right| \in \left[(j-1)\sqrt{n}, j\sqrt{n} \right) \right\}.$$

By Azuma's inequality, a random input drawn from $\mathcal{D}$ will with high probability be in Δ_j for small j:

$$\Pr_{X \sim \mathcal{D}} [X \in \Delta_j] \leq 2 \exp\left(-\frac{(j-1)^2}{2} \right). \tag{13}$$

In particular, for a random input $X \sim \mathcal{D}$, the expected value of the j such that $X \in \Delta_j$ is small:

$$\sum_{j=1}^{\infty} j \cdot \Pr_{\mathcal{D}} [\Delta_j] < 2. \tag{14}$$

Similarly, we can also bound higher moments of this function. Since the distribution over the Hamming weight of X is subgaussian with standard deviation $O(\sqrt{n})$, we also have:

$$\sum_{j=1}^{\infty} j^2 \cdot \Pr_{\mathcal{D}}[\Delta_j] = O(1). \tag{15}$$

We use $(\mathcal{D}|\Delta_j)$ to denote the distribution $\mathcal{D}$ conditioned on the event Δ_j (and similarly for $\mathcal{D}_0$). We proceed with a sequence of technical propositions, which will be used to bound the KL-divergence between $\mathcal{D}$ and $\mathcal{D}_0$.

Proposition 46 *Let Δ_j be as defined above. For every $j \geq 1$, every $X \in \Delta_j$, and every $b \in \{0,1\}$, we have:*

$$\left| \ln \frac{\Pr_{Y \sim \mathcal{D}}[u_t = b \mid Y]}{\Pr[u_t = b \mid X]} \right| \leq \frac{2j+5}{\sqrt{n}\mu}.$$

Proof. The intuition is that the probability that $u_t = b$ (by $\mathcal{D}$) is dominated by the probability that this event occurs for inputs whose Hamming weights are close to the expectation. By the differential privacy of the Laplace noise mechanism, the log-ratio of probabilities for inputs close to the expectation and inputs in Δ_j is upper-bounded by $\frac{j}{\sqrt{n}\mu}$ in magnitude. We show one direction (an upper bound); the lower bound follows similarly:

$$\begin{aligned}
\frac{\Pr_{Y \sim \mathcal{D}}[u_t = b \mid Y]}{\Pr[u_t = b \mid X]} &= \sum_{k=1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot \frac{\Pr_{\mathcal{D}}[u_t = b \mid \Delta_k]}{\Pr[u_t = b \mid X]} \\
&\leq \sum_{k=1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k+2j}{\sqrt{n}\mu}} \\
&= e^{\frac{2j}{\sqrt{n}\mu}} \left(\sum_{k=1}^{\sqrt{n}\mu} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k}{\sqrt{n}\mu}} + \sum_{k=\sqrt{n}\mu+1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k}{\sqrt{n}\mu}} \right) \\
&\leq e^{\frac{2j}{\sqrt{n}\mu}} \left(\sum_{k=1}^{\sqrt{n}\mu} \Pr_{\mathcal{D}}[\Delta_k] \cdot \left(1 + \frac{2k}{\sqrt{n}\mu}\right) + \sum_{k=\sqrt{n}\mu+1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k}{\sqrt{n}\mu}} \right) \\
&\leq e^{\frac{2j}{\sqrt{n}\mu}} \left(1 + \frac{2}{\sqrt{n}\mu} \sum_{k=1}^{\sqrt{n}\mu} \Pr_{\mathcal{D}}[\Delta_k] k + \sum_{k=\sqrt{n}\mu+1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k}{\sqrt{n}\mu}} \right) \\
&\leq e^{\frac{2j}{\sqrt{n}\mu}} \left(1 + \frac{4}{\sqrt{n}\mu} + \sum_{k=\sqrt{n}\mu+1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{\frac{k}{\sqrt{n}\mu}} \right) \\
&< e^{\frac{2j}{\sqrt{n}\mu}} \left(1 + \frac{5}{\sqrt{n}\mu} \right).
\end{aligned}$$

Here the second line follows from the differential privacy of the Laplace noise mechanism, as well as the fact that the Hamming distance between inputs $Y \in \Delta_k$ and $X \in \Delta_j$ is at most $(k+2j)\sqrt{n}$. The second-to-last line uses equation (14), while the final line uses equation (13), and can be seen

as follows:

$$\begin{aligned}\sum_{k=\sqrt{n}\mu+1}^{\sqrt{n}} \Pr_{\mathcal{D}}[\Delta_k] \cdot e^{k/(\sqrt{n}\mu)} &\leq \sum_{k=\sqrt{n}\mu}^{\sqrt{n}} \exp\left(\frac{-k^2}{2} + \frac{k+1}{\sqrt{n}\mu}\right) \\ &< 4e^{-n\mu^2/2} \\ &< \frac{1}{\sqrt{n}\mu},\end{aligned}$$

where the last two inequalities hold so long as $\sqrt{n}\mu$ is a sufficiently large constant. ■

Proposition 47 *Let Δ_j be as defined above. For every $j \geq 1$, every $X \in \Delta_j$, and every input X' that differs from X in a single coordinate, we have*

$$\left|\ln \frac{\Pr[u_t = 0 \mid X]}{\Pr[u_t = 0 \mid X']}\right| \leq \frac{4 \min\left\{1, e^{\frac{2j+5}{\sqrt{n}\mu}} \cdot \lambda_t\right\}}{n\mu}.$$

Proof. First, since we add Laplace noise of magnitude $n\mu$ before checking for an update, for *every* pair of adjacent inputs $X, X' \in \{0,1\}^n$, the log-ratio between the probabilities of $u_t = 0$ is at most $\frac{1}{n\mu}$. When the probability of an update is smaller, we can improve this bound as follows. Define q_t to be the probability of an update (i.e., $u_t = 1$) given the input X. By Proposition 46, we have $q_t \leq e^{\frac{2j+5}{\sqrt{n}\mu}} \cdot \lambda_t$.

Take the count on X to be $k = \sum_i X_i$. An update is triggered when the difference between the noisy count and $\mathrm{Tr}(E_t\sigma_{t-1})$ is too large—or equivalently, when the noisy count passes a threshold $h > k+1$.[18] Thus, $q_t = \Pr[k + \eta > h]$. Similarly, the probability q'_t of an update on X' is $\Pr[k + \eta + 1 > h]$. (The case where the count on X' is smaller than on X is handled similarly.) By the definition of the Laplace distribution, these probabilities are given by:

$$\begin{aligned}q_t &= \frac{1}{2}\exp\left(-\frac{h-k}{n\mu}\right), \\ q'_t &= \frac{1}{2}\exp\left(-\frac{h-k-1}{n\mu}\right)\end{aligned}$$

[18] This is without loss of generality: the case $h < k-1$ can be handled similarly. The case where $k \in [h-1, h+1]$ cannot occur because then λ_t would be much larger than say $\frac{1}{100}$, whereas we assumed λ_t was sufficiently small.

Now by standard manipulations we get:

$$\begin{aligned}
\frac{\Pr[u_t = 0 \mid X]}{\Pr[u_t = 0 \mid X']} &= \frac{1-q_t}{1-q'_t} \\
&= \frac{1 - \frac{1}{2}\exp\left(-\frac{h-k}{n\mu}\right)}{1 - \frac{1}{2}\exp\left(-\frac{h-k-1}{n\mu}\right)} \\
&= 1 + \frac{\exp\left(-\frac{h-k-1}{n\mu}\right) - \exp\left(-\frac{h-k}{n\mu}\right)}{2 - \exp\left(-\frac{h-k-1}{n\mu}\right)} \\
&\leq 1 + \exp\left(-\frac{h-k}{n\mu}\right)\left(\exp\left(\frac{1}{n\mu}\right) - 1\right) \\
&= 1 + 2q_t\left(\exp\left(\frac{1}{n\mu}\right) - 1\right) \\
&\leq \exp\left(\frac{4q_t}{n\mu}\right).
\end{aligned}$$

Here the last line uses the fact that $n\mu$ is a sufficiently large constant. (Note also that, in the case we're analyzing, the ratio of probabilities is larger than 1, so we only need to prove an *upper* bound.) Proposition 47 follows, recalling that by its conditions $q_t \leq e^{\frac{2j+5}{\sqrt{n}\mu}} \cdot \lambda_t$. ■

Proposition 48 *Let $\mathcal{D}$, $\mathcal{D}_0$ and Δ_j be as defined above. Then for every $j \geq 1$:*

$$\mathrm{KL}\left((\mathcal{D}|\Delta_j), (\mathcal{D}_0|\Delta_j)\right) \leq \frac{\min\left\{2, 32e^{\frac{4j+10}{\sqrt{n}\mu}} \cdot \lambda_t^2\right\}}{n\mu^2},$$

$$\mathrm{KL}\left((\mathcal{D}_0|\Delta_j), (\mathcal{D}|\Delta_j)\right) \leq \frac{\min\left\{2, 32e^{\frac{4j+10}{\sqrt{n}\mu}} \cdot \lambda_t^2\right\}}{n\mu^2}.$$

Proof. We employ a variant of the proof of Lemma 30. We spell out the bound in the first direction, the second direction follows similarly. Recall that $(\mathcal{D}|\Delta_j)$ is the product distribution $\mathcal{D}$, conditioned on the event Δ_j (the difference between the Hamming weight of X and its expectation is in the interval $[(j-1)\sqrt{n}, j\sqrt{n})$). We can sample an input $X = (x_1, \ldots, x_n)$ from this distribution by sampling x_1 from the marginal distribution over the first entry of $\mathcal{D}$ conditioned on Δ_j, then drawing x_2 from the marginal distribution over the second entry, conditioned on x_1 and Δ_j, and so on up to x_n. Call the i^{th} distribution $\mathcal{S}_i$; note that $\mathcal{S}_i$ depends on $x_1, \ldots, x_{i-1}$ (and on Δ_j). Similarly, we can also consider a conditional distribution $(\mathcal{D}_0|\Delta_j)$, where we condition both on $u_t = 0$ (no update) and on the event Δ_j occurring. We can sample from this second distribution by first drawing x_1 from the marginal distribution over the first entry conditioned on $u_t = 0$ and on Δ_j, then drawing x_2 from the marginal distribution over the second entry conditioned on no update, on Δ_j, and on x_1, and so on up to x_n. Call the i^{th} distribution in this second process $\mathcal{T}_i$; note that $\mathcal{T}_i$ depends on $x_1, \ldots, x_{i-1}$ (as well as on the set Δ_j and the event $u_t = 0$). The marginal distributions $\mathcal{S}_i$ and $\mathcal{T}_i$ are over $\{0, 1\}$.

We note that for any setting of the first $i-1$ variables, the supports of the random variables $\mathcal{S}_i$ and $\mathcal{T}_i$ are identical: a given prefix might make the event Δ_j impossible for a certain fixing of the

i^{th} variable, but in this case the forbidden fixing has weight 0 both in $\mathcal{S}_i$ and in $\mathcal{T}_i$. By Proposition 47 and by Bayes' rule, for every $i \in \{1, \ldots, n\}$, for every setting $x_1, \ldots, x_{i-1} \in \{0,1\}$ for the first $i-1$ input coordinates, and for every value $v \in \{0,1\}$ such that v has nonzero probability by $\mathcal{S}_i$, the magnitude of the log-ratio between v's probabilities under $\mathcal{S}_i$ and under $\mathcal{T}_i$ is bounded as follows:

$$\left|\ln \frac{\Pr_{\mathcal{S}_i}[v]}{\Pr_{\mathcal{T}_i}[v]}\right| = \left|\ln \frac{\Pr[u_t = 0 \mid x_1, \ldots, x_{i-1}, G]}{\Pr[u_t = 0 \mid x_1, \ldots, x_{i-1}, X_i = v, G]}\right| \leq \frac{\min\left\{1, 4e^{\frac{2j+5}{\sqrt{n}\mu}} \cdot \lambda_t\right\}}{n\mu}$$

By Claim 29, this means that the *expected* log-ratio between $\Pr_{\mathcal{S}_i}[v]$ and $\Pr_{\mathcal{T}_i}[v]$, with respect to x_i drawn from $\mathcal{S}_i$, is upper-bounded by

$$2\left(\frac{\min\left\{1, 4e^{\frac{2j+5}{\sqrt{n}\mu}} \cdot \lambda_t\right\}}{n\mu}\right)^2.$$

As in the proof of Lemma 30, we conclude:

$$\mathrm{KL}\left((\mathcal{D}|\Delta_j), (\mathcal{D}_0|\Delta_j)\right) = \sum_{i=1}^{n} \mathrm{E}\left[\mathrm{KL}\left(\mathcal{S}_i, \mathcal{T}_i\right)\right] \leq \frac{2\min\left\{1, 16e^{\frac{4j+10}{\sqrt{n}\mu}} \cdot \lambda_t^2\right\}}{n\mu^2}$$

■

Proposition 49 *Let $\mathcal{D}$, $\mathcal{D}_0$ and Δ_j be as defined above. Partition the line $[1, \sqrt{n}]$ into the following three segments:*

$$J_1 = \left[1, \frac{\sqrt{n}\mu - 5}{2}\right)$$
$$J_2 = \left[\frac{\sqrt{n}\mu - 5}{2}, \frac{\sqrt{n}\mu(\ln(\frac{1}{\lambda_t}) - 1) - 5}{2}\right)$$
$$J_3 = \left[\frac{\sqrt{n}\mu(\ln(\frac{1}{\lambda_t}) - 1) - 5}{2}, \sqrt{n}\right].$$

Then the following hold:

- *For every integer $j \in J_1$:*
$$\left|\ln \frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]}\right| \leq \frac{\lambda_t(4j+10)}{\sqrt{n}\mu}.$$
- *For every integer $j \in J_2$:*
$$\left|\ln \frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]}\right| \leq 2\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}.$$
- *For every integer $j \in J_3$:*
$$\left|\ln \frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]}\right| \leq \frac{2j+5}{\sqrt{n}\mu}.$$

Proof. First, by Bayes' rule, for every j we have:

$$\frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]} = \frac{\Pr_{\mathcal{D}}[u_t = 0]}{\Pr_{\mathcal{D}}[u_t = 0 \mid \Delta_j]}.$$

Further, by Proposition 46 we have that for every integer $j \geq 1$:

$$\left| \ln \frac{\Pr_{\mathcal{D}}[u_t = b]}{\Pr_{\mathcal{D}}[u_t = b \mid \Delta_j]} \right| \leq \frac{2j+5}{\sqrt{n}\mu}. \tag{16}$$

(The Proposition asserts this for every $X \in \Delta_j$; the claim when conditioning on Δ_j follows by a standard argument.)

Case analysis. We proceed to analyze each of the cases separately, beginning with the case $j \in J_1$. Recall that $\Pr_{\mathcal{D}}[u_t = 1] = \lambda_t$. By equation (16), the probability of $u_t = 1$ under Δ_j can differ from this by at most an $e^{\frac{(2j+5)\sqrt{n}}{n\mu}}$ multiplicative factor. We conclude that:

$$\begin{aligned}
\frac{\Pr_{\mathcal{D}}[u_t = 0]}{\Pr_{\mathcal{D}}[u_t = 0 \mid \Delta_j]} &\leq \frac{1-\lambda_t}{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}} \\
&= \frac{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}}{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}} + \frac{\lambda_t \left(e^{\frac{2j+5}{\sqrt{n}\mu}} - 1 \right)}{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}} \\
&\leq 1 + \frac{\lambda_t \left(e^{\frac{2j+5}{\sqrt{n}\mu}} - 1 \right)}{1/2} \\
&\leq e^{\frac{\lambda_t(4j+10)}{\sqrt{n}\mu}}.
\end{aligned}$$

Here the second-to-last line holds because for this range of j we have $2j + 5 \leq \sqrt{n}\mu$, and thus $\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}} < \frac{1}{2}$. The last line holds because for the same range of j we have

$$e^{\frac{2j+5}{\sqrt{n}\mu}} - 1 \leq \frac{4j+10}{\sqrt{n}\mu}.$$

To conclude the analysis of the first case, observe that for similar reasons also in the other direction we have:

$$\frac{\Pr_{\mathcal{D}}[u_t = 0]}{\Pr_{\mathcal{D}}[u_t = 0 \mid \Delta_j]} \geq e^{\frac{-\lambda_t(4j+10)}{\sqrt{n}\mu}}.$$

For the second case, $j \in J_2$, we have $\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}} \leq \frac{1}{e}$ and thus:

$$\begin{aligned}
\frac{\Pr_{\mathcal{D}}[u_t = 0]}{\Pr_{\mathcal{D}}[u_t = 0 \mid \Delta_j]} &\leq \frac{1-\lambda_t}{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}} \\
&\leq \frac{1}{1-\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}} \\
&\leq e^{2\lambda_t e^{(2j+5)/(\sqrt{n}\mu)}}.
\end{aligned}$$

Here the last line holds because for all $z \in (0, e^{-1}]$ we have $\frac{1}{1-z} \leq e^{-2z}$. In the other direction:

$$\begin{aligned}\frac{\Pr_{\mathcal{D}}[u_t = 0]}{\Pr_{\mathcal{D}}[u_t = 0 \mid \Delta_j]} &\geq \Pr_{\mathcal{D}}[u_t = 0] \\ &= 1 - \lambda_t \\ &\geq e^{-2\lambda_t}.\end{aligned}$$

The third case, $j \in J_3$, follows immediately from equation (16) (or Proposition 49), which holds for every possible value of j.

Bounding the KL-divergence. We now proceed to bound the KL-divergence between $\mathcal{D}$ and $\mathcal{D}_0$

$$\begin{aligned}\mathrm{KL}\left(\mathcal{D}, \mathcal{D}_0\right) &= \sum_{X \in \{0,1\}^n} \Pr_{\mathcal{D}}[X] \ln \frac{\Pr_{\mathcal{D}}[X]}{\Pr_{\mathcal{D}_0}[X]} \\ &= \sum_j \left(\sum_{X \in \Delta_j} \Pr_{\mathcal{D}}[X] \ln \frac{\Pr_{\mathcal{D}}[X]}{\Pr_{\mathcal{D}_0}[X]} \right) \\ &= \sum_j \left(\sum_{X \in \Delta_j} \Pr_{\mathcal{D}}[\Delta_j] \Pr_{(\mathcal{D}|\Delta_j)}[X] \ln \frac{\Pr_{\mathcal{D}}[\Delta_j] \Pr_{(\mathcal{D}|\Delta_j)}[X]}{\Pr_{\mathcal{D}_0}[\Delta_j] \Pr_{(\mathcal{D}_0|\Delta_j)}[X]} \right) \\ &= \sum_j \left(\Pr_{\mathcal{D}}[\Delta_j] \ln \frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]} + \Pr_{\mathcal{D}}[\Delta_j] \sum_{X \in \Delta_j} \Pr_{(\mathcal{D}|\Delta_j)}[X] \ln \frac{\Pr_{(\mathcal{D}|\Delta_j)}[X]}{\Pr_{(\mathcal{D}_0|\Delta_j)}[X]} \right) \\ &= \sum_j \left(\Pr_{\mathcal{D}}[\Delta_j] \ln \frac{\Pr_{\mathcal{D}}[\Delta_j]}{\Pr_{\mathcal{D}_0}[\Delta_j]} + \Pr_{\mathcal{D}}[\Delta_j] \,\mathrm{KL}\left((\mathcal{D}|\Delta_j), (\mathcal{D}_0|\Delta_j)\right) \right)\end{aligned}$$

And similarly:

$$\mathrm{KL}\left(\mathcal{D}_0, \mathcal{D}\right) = \sum_j \left(\Pr_{\mathcal{D}_0}[\Delta_j] \ln \frac{\Pr_{\mathcal{D}_0}[\Delta_j]}{\Pr_{\mathcal{D}}[\Delta_j]} + \Pr_{\mathcal{D}_0}[\Delta_j] \,\mathrm{KL}\left((\mathcal{D}_0|\Delta_j), (\mathcal{D}|\Delta_j)\right) \right)$$

Using the nonnegativity of KL-divergence, together with the bound in Proposition 48, we conclude that:

$$\begin{aligned}\mathrm{KL}\left(\mathcal{D}, \mathcal{D}_0\right) \leq& \,\mathrm{KL}\left(\mathcal{D}, \mathcal{D}_0\right) + \mathrm{KL}\left(\mathcal{D}_0, \mathcal{D}\right) \\ =& \sum_j (\Pr_{\mathcal{D}_0}[\Delta_j] - \Pr_{\mathcal{D}}[\Delta_j]) \ln \frac{\Pr_{\mathcal{D}_0}[\Delta_j]}{\Pr_{\mathcal{D}}[\Delta_j]} \\ &+ \sum_j \left(\Pr_{\mathcal{D}}[\Delta_j] + \Pr_{\mathcal{D}_0}[\Delta_j] \right) \frac{2 \min\left\{1, 16 e^{\frac{4j+10}{\sqrt{n}\mu}} \cdot \lambda_t^2\right\}}{n\mu^2}.\end{aligned}$$

Below, we show that each of these two sums is bounded by $O\left(\frac{\lambda_t^2}{n\mu^2}\right)$. We conclude that

$$\left\| |\psi\rangle\langle\psi| - |\psi_{\mathrm{no}}\rangle\langle\psi_{\mathrm{no}}| \right\|_{\mathrm{tr}} \leq \sqrt{2 \,\mathrm{KL}\left(\mathcal{D}, \mathcal{D}_0\right)} = O\left(\frac{\lambda_t}{\sqrt{n}\mu}\right),$$

which completes the proof of Claim 41. ■

Bounding the first sum. We divide the sum over j into the three segments defined in Proposition 49, and use the bound on the log-ratio to bound the sum over each of the segments. Starting with the first segment J_1:

$$\begin{aligned}
\sum_{j\in J_1}(\Pr_{\mathcal{D}_0}[\Delta_j]-\Pr_{\mathcal{D}}[\Delta_j])\ln\frac{\Pr_{\mathcal{D}_0}[\Delta_j]}{\Pr_{\mathcal{D}}[\Delta_j]} &\le \sum_{j\in J_1}\Pr_{\mathcal{D}}[\Delta_j]\left(e^{\frac{\lambda_t(4j+10)}{\sqrt{n}\mu}}-1\right)\frac{\lambda_t(4j+10)}{\sqrt{n}\mu}\\
&\le 8\sum_{j\in J_1}\Pr_{\mathcal{D}}[\Delta_j]\frac{\lambda_t^2(4j+10)^2}{n\mu^2}\\
&= 8\frac{\lambda_t^2}{n\mu^2}\sum_{j\in J_1}\Pr_{\mathcal{D}}[\Delta_j](4j+10)^2\\
&= \Theta\left(\frac{\lambda_t^2}{n\mu^2}\right).
\end{aligned}$$

Here the first line uses Proposition 49; the second line uses the fact that $4j+10\le 2\sqrt{n}\mu$ for all $j\in J_1$; and the last line uses a moment bound on the distribution of j, namely inequality (15).

For the segment J_2 we have:

$$\begin{aligned}
\sum_{j\in J_2}(\Pr_{\mathcal{D}_0}[\Delta_j]-\Pr_{\mathcal{D}}[\Delta_j])\ln\frac{\Pr_{\mathcal{D}_0}[\Delta_j]}{\Pr_{\mathcal{D}}[\Delta_j]} &\le \sum_{j\in J_2}\Pr_{\mathcal{D}}[\Delta_j]\left(e^{2\lambda_t e^{(2j+5)/(\sqrt{n}\mu)}}-1\right)2\lambda_t e^{\frac{2j+5}{\sqrt{n}\mu}}\\
&\le 8\lambda_t^2\sum_{j\in J_2}\Pr_{\mathcal{D}}[\Delta_j]e^{\frac{4j+10}{\sqrt{n}\mu}}\\
&\le 16\lambda_t^2\sum_{j\in J_2}\exp\left(\frac{4j+10}{\sqrt{n}\mu}-\frac{(j-1)^2}{2}\right)\\
&= 16\lambda_t^2\sum_{j\in J_2}\exp\left(-\Theta\left(j^2\right)\right)\\
&= 16\lambda_t^2\exp\left(-\Theta\left(n\mu^2\right)\right)\\
&= O\left(\frac{\lambda_t^2}{n\mu^2}\right).
\end{aligned}$$

Here the first line uses Proposition 49, the second uses the fact that $\frac{2j+5}{\sqrt{n}\mu}\le\ln(\frac{1}{\lambda_t})-1$ for all $j\in J_2$; the third uses inequality (13) (concentration of j); and the fourth and fifth use the facts that $j=\Omega(\sqrt{n}\mu)$ for all $j\in J_2$ and that$\sqrt{n}\mu$ is a sufficiently large constant.

For the segment J_3 we have:

$$\begin{aligned}\sum_{j\in J_3}(\Pr_{\mathcal{D}_0}[\Delta_j]-\Pr_{\mathcal{D}}[\Delta_j])\ln\frac{\Pr_{\mathcal{D}_0}[\Delta_j]}{\Pr_{\mathcal{D}}[\Delta_j]} &\le \sum_{j\in J_3}\Pr_{\mathcal{D}}[\Delta_j]\left(e^{\frac{2j+5}{\sqrt{n}\mu}}-1\right)\frac{2j+5}{\sqrt{n}\mu}\\ &\le 2\sum_{j\in J_3}e^{-(j-1)^2/2}\left(e^{\frac{2j+5}{\sqrt{n}\mu}}-1\right)\frac{2j+5}{\sqrt{n}\mu}\\ &= \sum_{j\in J_3}\exp\left(-\Theta(j^2)\right)\\ &= \exp\left(-\Theta(n\mu^2\ln^2\frac{1}{\lambda_t}\right)\\ &= O\left(\frac{\lambda_t^2}{n\mu^2}\right).\end{aligned}$$

Here the first line uses Proposition 49; the second uses inequality (13) (concentration of j); and the third and fourth use the facts that $j=\Omega(\sqrt{n}\mu\ln\frac{1}{\lambda_t})$ for all $j\in J_3$ and that $\sqrt{n}\mu$ is a sufficiently large constant.

Bounding the second sum. Similarly to the first sum, we divide the sum over j into the three segments defined in Proposition 49, and use the bound on the log-ratio to bound the sum over each of the segments. Starting with the first segment J_1:

$$\begin{aligned}\sum_{j\in J_1}\left(\Pr_{\mathcal{D}}[\Delta_j]+\Pr_{\mathcal{D}_0}[\Delta_j]\right)\frac{\min\left\{2,32e^{\frac{4j+10}{\sqrt{n}\mu}}\cdot\lambda_t^2\right\}}{n\mu^2} &\le \sum_{j\in J_1}\Pr_{\mathcal{D}}[\Delta_j]\left(1+e^{\frac{\lambda_t(4j+10)}{\sqrt{n}\mu}}\right)\frac{32e^{\frac{4j+10}{\sqrt{n}\mu}}\cdot\lambda_t^2}{n\mu^2}\\ &= \Theta\left(\frac{\lambda_t^2}{n\mu^2}\right).\end{aligned}$$

Here the first inequality uses Proposition 49, while the second uses the fact that $4j+10\le 2\sqrt{n}\mu$ for all $j\in J_1$.

For the segment J_2 we have:

$$\begin{aligned}\sum_{j\in J_2}\left(\Pr_{\mathcal{D}}[\Delta_j]+\Pr_{\mathcal{D}_0}[\Delta_j]\right)\frac{\min\left\{2,32e^{\frac{4j+10}{\sqrt{n}\mu}}\cdot\lambda_t^2\right\}}{n\mu^2} &\le \sum_{j\in J_2}\Pr_{\mathcal{D}}[\Delta_j]\left(1+e^{2\lambda_t e^{(2j+5)/(\sqrt{n}\mu)}}\right)\frac{32e^{\frac{4j+10}{\sqrt{n}\mu}}\cdot\lambda_t^2}{n\mu^2}\\ &\le \frac{128\lambda_t^2}{n\mu^2}\sum_{j\in J_2}\Pr_{\mathcal{D}}[\Delta_j]e^{\frac{4j+10}{\sqrt{n}\mu}}\\ &\le \frac{256\lambda_t^2}{n\mu^2}\sum_{j\in J_2}\exp\left(\frac{4j+10}{\sqrt{n}\mu}-\frac{(j-1)^2}{2}\right)\\ &= O\left(\frac{\lambda_t^2}{n\mu^2}\right).\end{aligned}$$

Here the first line uses Proposition 49; the second uses the fact that $\frac{2j+5}{\sqrt{n}\mu}\le\frac{1}{e\lambda_t}$ for all $j\in J_2$; the third uses equation (13); and the last uses the facts that $j=\Omega(\sqrt{n}\mu)$ for all $j\in J_2$ and that $\sqrt{n}\mu$ is a sufficiently large constant.

Finally, for the segment J_3 we have:

$$\begin{aligned}\sum_{j\in J_3}\left(\Pr_{\mathcal{D}}[\Delta_j]+\Pr_{\mathcal{D}_0}[\Delta_j]\right)\frac{\min\left\{2,32e^{\frac{4j+10}{\sqrt{n}\mu}}\cdot\lambda_t^2\right\}}{n\mu^2} &\leq \sum_{j\in J_3}\Pr_{\mathcal{D}}[\Delta_j]\left(1+e^{\frac{2j+5}{\sqrt{n}\mu}}\right)\frac{2}{n\mu^2}\\ &\leq \frac{4}{n\mu^2}\sum_{j\in J_3}\exp\left(\frac{2j+5}{\sqrt{n}\mu}-\frac{(j-1)^2}{2}\right)\\ &= \frac{4}{n\mu^2}\exp\left(-\Theta\left(n\mu^2\ln^2\frac{1}{\lambda_t}\right)\right)\\ &= O\left(\frac{\lambda_t^2}{n\mu^2}\right).\end{aligned}$$

Here the first line uses Proposition 49; the second uses equation (13); and the third uses the facts that $j=\Omega\left(\sqrt{n}\mu\ln\frac{1}{\lambda_t}\right)$ for all $j\in J_3$ and that $\sqrt{n}\mu$ is a sufficiently large constant. ■

6.3 Lower Bounds for Shadow Tomography

To recap, the QPMW algorithm lets us do shadow tomography on a d-dimensional state ρ, with respect to two-outcome measurements $E_1,\ldots,E_m$ and with accuracy $\pm\varepsilon$, in a way that moreover is online and gentle, by measuring $O\left((\log m)^2(\log d)^2/\varepsilon^8\right)$ copies of ρ. How close to optimal is this upper bound?

The only known general lower bound for shadow tomography, due to Aaronson [6], says that $\Omega\left(\frac{\min\left\{d^2,\log m\right\}}{\varepsilon^2}\right)$ copies of ρ are needed, for information-theoretic reasons. Aaronson [6] also shows that, in the special case where the states and measurements are entirely classical, $\Theta\left(\frac{\min\{d,\log m\}}{\varepsilon^2}\right)$ copies are necessary and sufficient.[19] In the general, quantum setting, it remains open whether there could exist a shadow tomography procedure that used only $(\log m)^{O(1)}$ copies, independent of the dimension d.

In this section, we won't resolve that problem. However, as yet another application of our connection between DP and gentleness, we'll observe a lower bound on the sample complexity of *gentle* shadow tomography, which applies even to offline algorithms—i.e., ones that see all the measurements in advance. And conversely, by using the connection to adaptive data analysis, we'll use known results in that setting to give a lower bound for *online* shadow tomography, which applies even to non-gentle algorithms.

We stress that, while these lower bounds use nontrivial recent results, they have nothing to do with quantum mechanics: all of them apply even to the "classical special case" of shadow tomography, wherein the input consists of i.i.d. samples from a single distribution and the "measurements" are all in the computational basis.

Gentle shadow tomography. The first result we state is a lower bound for *gentle* shadow tomography, even in the offline setting:

[19] The original conference version of [6] proved only a weaker lower bound: namely, $\Omega\left(\frac{\log m}{\varepsilon^2}\right)$ when d can be arbitrarily large (including for the classical special case). However, the most recent arXiv version includes the stated bounds, the ones that explicitly incorporate dependence on the dimension d.

Theorem 50 (Lower Bound for Gentle Shadow Tomography) *Any shadow tomography procedure that is $\left(\alpha, \frac{1}{n^{1+\tau}}\right)$-gentle for a constant $\tau > 0$ on all product states, and is also ε-accurate on states of the form $\rho^{\otimes n}$, requires*

$$n = \widetilde{\Omega}\left(\frac{(\log m)\sqrt{\log d}}{\varepsilon^2 \alpha}\right)$$

samples.

In other words, as long as we insist that our shadow tomography procedure be (α, δ)-gentle for small δ—with gentleness applying for all product states, as usual in this paper—the sample complexity of the QPMW algorithm is optimal up to a polynomial factor.

We'll deduce Theorem 50 as a corollary of the following result of Bun, Ullman, and Vadhan [16]:

Theorem 51 (Bun et al. [16]) *For all m, n, d, there exist m Boolean functions $f_1, \ldots, f_m : [d] \to \{0,1\}$, such that no $\left(\gamma, \frac{1}{10n}\right)$-DP algorithm can, for all databases $X = (x_1, \ldots, x_n) \in [d]^n$, estimate $\mathrm{E}_{j \in [n]}\left[f_i\left(x_j\right)\right]$ to within additive error $\pm\varepsilon$, for every $i \in [m]$ and with success probability at least $2/3$, unless*

$$n = \widetilde{\Omega}\left(\frac{(\log m)\sqrt{\log d}}{\varepsilon^2 \gamma}\right).$$

The proof of Theorem 51 uses so-called *fingerprinting codes* to construct the functions $f_1, \ldots, f_m$. We omit the details; see for example Vadhan [44, Section 5.3] for further discussion of this technique.

Recall Lemma 28, which said that any measurement that is (α, δ)-gentle on product states is also $\left(\ln\left(\frac{1+4\alpha}{1-4\alpha}\right), \delta\right)$-DP on product states. In the classical special case, the latter simply means $\left(\ln\left(\frac{1+4\alpha}{1-4\alpha}\right), \delta\right)$-DP in the usual sense. By just combining this implication with Theorem 51, we immediately obtain a lower bound on the sample complexity of *some* form of gentle shadow tomography, even in the classical special case. However, there is still a difficulty. Namely, the lower bound that we get will apply only to shadow tomography algorithms that remain accurate in what we call the *diverse-state setting*. This is the setting where the algorithm is given a sample from a product distribution $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$—or in the quantum case, a product state $\rho_1 \otimes \cdots \otimes \rho_n$—and its goal is to estimate the acceptance probability of each of the two-outcome measurements $E_1, \ldots, E_m$ on the average state

$$\widehat{\rho} := \frac{\rho_1 + \cdots + \rho_n}{n}.$$

By contrast, we defined shadow tomography for what we call the *identical-state setting*: that is, the setting where we're additionally promised that $\rho_1 = \cdots = \rho_n$, so that the input state has the special form $\rho^{\otimes n}$. All of the shadow tomography procedures that we know, including QPMW, are accurate even in the more general diverse-state setting. But it's not obvious that *lower bounds* in the diverse-state setting carry over to the identical-state setting, so there is still a gap to close. We close the gap using the simple claim below, which translates accuracy in the identical-state setting to accuracy in the diverse-state setting, with only a small loss in the differential privacy parameters.[20]

[20] We note that it might be possible to obtain a lower bound similar to that of Theorem 51 that directly applies to the identical-state setting (see, e.g., Steinke and Ullman [41, Corollary 15]). Still, the transformation we outline incurs only a small loss in the parameters, and works more generally.

Claim 52 *Fix a data universe $[d]$, functions $f_1, \ldots, f_m : [d] \to \{0,1\}$, and a database size n. Let $\mathcal{A}$ be a classical algorithm that's (α, δ)-DP in the usual classical sense, and satisfies the following accuracy guarantee: for any distribution $\mathcal{D}$ over $[d]$, with all but β probability over X drawn from $\mathcal{D}^n$ (and the algorithm's coins), $\mathcal{A}$'s answers are all within $\pm\varepsilon$ of the correct answers $\mathrm{E}_{j\in[n]}\left[f_i\left(X_j\right)\right]$.*

Then there exists another algorithm $\mathcal{B}$, which runs in $O(|X|)$ time using a single oracle call to $\mathcal{A}$, such that $\mathcal{B}$ is $(\frac{3\alpha\log n}{\log\log n}, \delta\cdot n^{o(1)} + \frac{1}{n^2})$-DP, and for any fixed database $X \in [d]^n$, with all but β probability over the $\mathcal{B}$'s coins, $\mathcal{B}$'s answers are all within ε of the correct answers $\mathrm{E}_{j\in[n]}\left[f_i\left(X_j\right)\right]$.

Proof. Given an input database X, the algorithm $\mathcal{B}$ operates by taking n i.i.d. samples, with replacement, from the distribution $\mathcal{D}_X$ that is uniform over the entries of X (a distribution whose support has size at most n). It then runs $\mathcal{A}$ on the resulting database X' and outputs the results.

Accuracy follows because we are running $\mathcal{A}$ on a sample from $\mathcal{D}_X^n$, so with all but β probability over the samples and $\mathcal{A}$'s coins, the answers will all be within $\pm\varepsilon$ of the correct expectations over $\mathcal{D}_X$, which are the correct answers on the database X.

For privacy, fix adjacent databases X and Y that differ only in the i^{th} entry. For fixed coins used to choose i.i.d. samples, let X' and Y' be the databases produced by sampling from X or from Y respectively. All entries in X' and Y' will be identical, except those that are copies of the i^{th} entry. By a balls and bins argument, with all but $1/n^2$ probability, the number of copies of the i^{th} entry is at most $\frac{3\log n}{\log\log n}$. Whenever this is the case, the group privacy guarantees that follow from the differential privacy of $\mathcal{A}$ imply that the probability of any event differs by at most a $\frac{3\alpha\log n}{\log\log n}$ multiplicative factor and a $\delta\cdot n^{o(1)}$ additive error. ■

We can now complete the proof of Theorem 50.

Proof of Theorem 50. Let $\mathcal{P}$ be a shadow tomography procedure that is $\left(\alpha, \frac{1}{n^{1+\tau}}\right)$-gentle on product states $\rho_1\otimes\cdots\otimes\rho_n$, for small α and fixed $\tau > 0$. By Lemma 28, this $\mathcal{P}$ is also (γ,δ)-DP on product states, for $\gamma = \ln\left(\frac{1+4\alpha}{1-4\alpha}\right) = O\left(\alpha\right)$.

Henceforth, we restrict attention to $\mathcal{P}$'s behavior on classical inputs $X \in [d]^n$. Here, being DP on product states simply reduces to the usual notion of DP.

Now suppose further that $\mathcal{P}$ is ε-accurate in the identical-state setting. Then by Claim 52, we can obtain a new classical procedure $\mathcal{P}'$ that is $\left(\gamma' = \frac{3\gamma\log n}{\log\log n}, \delta' o(1) = \delta + \frac{1}{n^2}\right)$-DP, and that moreover is ε-accurate for any given database $X \in [d]^n$. But this means that n must satisfy the bound of Theorem 51. We use here the fact that for $\delta = \frac{1}{n^{1+\tau}}$ we get $\delta' = o(\frac{1}{n})$. We note that the $O(\frac{\log n}{\log\log n})$ deterioration in the privacy guarantee of $\mathcal{P}'$ (compared to $\mathcal{P}$) is accounted for by the tilde in the $\widetilde{\Omega}$. ■

As noted above, Theorem 50 applies even to the "classical special case" of shadow tomography. In that special case, the Chernoff bound immediately implies a procedure with $O\left(\frac{\log m}{\varepsilon^2}\right)$ sample complexity. Thus, one implication of Theorem 50 is that such a procedure necessarily violates gentleness—where "gentleness," here, means a bound on the damage in variation distance caused by classical Bayesian updating.

Online shadow tomography. The second result we state is a lower bound for *online* shadow tomography, even without gentleness:

Theorem 53 (Lower Bound for Online Shadow Tomography [35]) *Any online shadow tomography procedure that is ε-accurate requires sample complexity*

$$n = \widetilde{\Omega}\left(\frac{\sqrt{\min\{m, \log d\}}}{\varepsilon}\right).$$

Combining Theorem 53 with the $\Omega\left(\frac{\min\{d, \log m\}}{\varepsilon^2}\right)$ lower bound of Aaronson [6], we can conclude that online shadow tomography requires

$$n = \widetilde{\Omega}\left(\frac{\log m + \sqrt{\log d}}{\varepsilon}\right)$$

copies of ρ unless $n = m^{\Omega(1)}$ or $n = d^{\Omega(1)}$. Hence QPMW achieves the optimal sample complexity for online shadow tomography up to polynomial factors.

Theorem 53 again has nothing to do with quantum mechanics, and follows immediately from known lower bounds for classical adaptive data analysis. There, an algorithm processes a collection of states that are drawn i.i.d. from an underlying distribution, and the goal is to provide accurate answers with respect to the underlying distribution—and in particular, to avoid overfitting to the specific sample. Adaptive data analysis is thus a special case of online shadow tomography in the identical-state setting. Hardt and Ullman [29] and Steinke and Ullman [40] showed sample complexity lower bounds and computational hardness results for this setting. Theorem 53 is a restatement, in our setting, of a recent result of Nissim et al. [35].

7 Computational Efficiency

So far, our results have been purely information-theoretic. When we talked, for example, about a gentle "implementation" of a measurement M, we were concerned only about whether such an implementation existed, not about its time complexity. Likewise, the QPMW procedure for shadow tomography was efficient in sample complexity, but we weren't concerned to bound its computation time.

Now, at last, we consider to what extent our constructions are (or can be made) computationally efficient. In Section 7.1, we'll explain why gentle measurements *can* be implemented in polynomial time, provided we have an efficient way to uncompute garbage, and we'll give several classes of examples where this can be done. Then, in Section 7.2, we'll use our results from Section 7.1 to examine the computational complexity of the QPMW procedure. Finally, in Section 7.3, we'll turn things around, and observe how gentle measurements like the ones in this paper, whether derived from DP algorithms or not, can be applied to the safe implementation of subroutines in quantum algorithms.

7.1 Efficiency of DP and Gentle Measurements

Let's start with Theorem 5, the connection between gentleness and DP. For part (1) of the theorem, namely that α-gentleness implies $O(\alpha)$-DP for small α, there's no issue of computational efficiency. This is because the very same measurement procedure that achieves α-gentleness also achieves $O(\alpha)$-DP—the latter being solely a property of the output probabilities, which has nothing to do with the post-measurement states.

On the other hand, for part (2) of the theorem, namely that ε-DP on product states implies $O(\varepsilon\sqrt{n})$-gentleness on product states for small ε (and product measurements), there *is* a computational issue. Namely: even if our original ε-DP measurement M could be implemented by a polynomial-size circuit, the proof of Theorem 5 might return an implementation of M that is $O(\varepsilon\sqrt{n})$-gentle but that does *not* correspond to any polynomial-size circuit. Yet, while this is a problem in principle, fortunately it turns out not to be a problem for any of the measurements that have concerned us in this paper, including the ones used in our shadow tomography procedure.

The potential computational issue occurs in the proof of Lemma 31. There, given a classical DP algorithm A, we needed to map the state

$$|\psi\rangle = \sum_{X \in [d]^n} \alpha_X |X\rangle$$

to

$$\sum_{X \in [d]^n : \Pr[X] > 0,\ y} \alpha_X |X\rangle \sqrt{\Pr[y|X]}\, |y\rangle,$$

where the y's are the possible outcomes of running A on the input database X. Assuming that A itself is computationally efficient, it's easy to prepare a state of the form

$$\sum_{X \in [d]^n : \Pr[X] > 0,\ y} \alpha_X |X\rangle \sqrt{\Pr[y|X]}\, |y\rangle\, |g_{X,y}\rangle,$$

where $|g_{X,y}\rangle$ is "garbage" entangled with the $|X\rangle$ and $|y\rangle$ registers (for example, the outcomes of coin flips made by A). The entire difficulty lies in uncomputing the $|g_{X,y}\rangle$ register. If we fail to uncompute, then the effect on $|\psi\rangle$ might no longer be gentle.

As we mentioned in Section 1.3, an equivalent way to say this is that our reduction from DP to gentleness preserves efficiency if, and only if, we have an efficient algorithm to "QSample" the output distribution of the DP algorithm A, meaning to prepare the superposition

$$|\phi_X\rangle := \sum_y \sqrt{\Pr[y|X]}\, |y\rangle$$

for a given input X. In practice, many fast sampling algorithms do give rise to fast QSampling algorithms, but this need not always be the case. Indeed, as pointed out by Aharonov and Ta-Shma [9] in 2003, if fast sampling always implied fast QSampling, then we'd immediately get polynomial-time quantum algorithms for graph isomorphism, breaking lattice-based cryptosystems, and all other problems in the class SZK (Statistical Zero Knowledge). Closely related to that, the collision lower bound of Aaronson [1] implies that, in the black-box setting, fast sampling does *not* imply fast QSampling.

But what about the specific measurements considered in this paper? Let's start with the following observation:

Proposition 54 (Efficient Implementation of L_σ) *There is an $O(n)$-size quantum circuit to implement L_σ, the Laplace noise measurement on n qubits, to $\frac{1}{\exp(n)}$ accuracy, so long as $\sigma = \exp(O(n))$.*

Proof. We simply use the procedure for implementing L_σ described in Section 1.1: the one where, given a superposition over $|X\rangle$'s, we first prepare a Laplace noise register

$$|\eta\rangle := \frac{1}{Z}\sum_{k=-C}^{C} e^{-|k|/(2\sigma)}\,|k\rangle\,,$$

for some cutoff $C = \exp(O(n))$ and normalization Z, then calculate $||X| + k\rangle$, and finally use $|X\rangle$ and $||X| + k\rangle$ together to uncompute the noise $|k\rangle$. What makes this work is that, in L_σ, the noise is entirely additive, and addition of integers is an easy operation to invert.

Also, as long as $\sigma = \exp(O(n))$, a cutoff of the form $C = \exp(O(n))$ suffices for exponential accuracy. Moreover, one can check that DP, and hence gentleness, still hold even after we impose the cutoff.

It remains only to verify that there are $O(n)$-size quantum circuits to add and subtract $O(n)$-bit integers, and to prepare $|\eta\rangle$. The one interesting part is preparing $|\eta\rangle$. Omitting normalization and restricting to $k \geq 0$ for simplicity, we observe that

$$\sum_{k=0}^{2^n-1} e^{-\gamma k}\,|k\rangle = \left(|0\rangle + e^{-\gamma 2^{n-1}}\,|1\rangle\right) \otimes \left(|0\rangle + e^{-\gamma 2^{n-2}}\,|1\rangle\right) \otimes \cdots \otimes \left(|0\rangle + e^{-\gamma}\,|1\rangle\right),$$

from which a linear-size circuit to prepare $|\eta\rangle$ follows. ■

Note that the algorithm from Proposition 54 is "maximally gentle," in the sense that for every possible state ρ of the n input registers (including non-product states), the *only* damage that running the algorithm causes to ρ, is the damage that necessarily results from learning the desired output.

We now prove a much more general result, though one that's formally incomparable to Proposition 54. We start with a trivial-seeming proposition.

Proposition 55 *Suppose we have two polynomial-time quantum algorithms: an algorithm A that, given a classical string X, prepares a state $|\zeta_X\rangle$, and an algorithm B that, for some $k = n^{O(1)}$, maps $|\zeta_X\rangle^{\otimes k}$ to $|\zeta_X\rangle^{\otimes k}\,|\phi_X\rangle$, to $\frac{1}{n^{O(1)}}$ accuracy. Then there's also a polynomial-time quantum algorithm Q that maps $|X\rangle$ to $|X\rangle\,|\phi_X\rangle$, to $\frac{1}{n^{O(1)}}$ accuracy.*

Proof. We first run A sequentially k times, to map $|X\rangle$ to $|X\rangle\,|\zeta_X\rangle^{\otimes k}$. We next run B, to map $|X\rangle\,|\zeta_X\rangle^{\otimes k}$ to $|X\rangle\,|\zeta_X\rangle^{\otimes k}\,|\phi_X\rangle$ (to $\frac{1}{n^{O(1)}}$ accuracy). Finally we run $A^\dagger$ sequentially k times, to map $|X\rangle\,|\zeta_X\rangle^{\otimes k}\,|\phi_X\rangle$ to $|X\rangle\,|\phi_X\rangle$. ■

Despite its simplicity, Proposition 55 lets us efficiently implement a large class of gentle measurements: namely, any gentle measurement that admits an efficient "two-part algorithm," wherein the first part prepares states $|\zeta_X\rangle$ (which might include unwanted garbage), and the second part maps the $|\zeta_X\rangle$ states to a desired output state $|\phi_X\rangle$ that—crucially—is nearly unentangled with the $|\zeta_X\rangle$'s, depending only on the original input X.

Let's give an example.

Theorem 56 (Fast QSampling of Sparse Distributions) *For each input X, suppose the state $|\phi_X\rangle$ has the form*

$$|\phi_X\rangle = \sum_{y \in S_X} \sqrt{\Pr[y|X]}\,|y\rangle\,,$$

where the support sets $S_X \subset \{0,1\}^m$ all satisfy $|S_X| \leq \ell$, for some $\ell = n^{O(1)}$ (i.e., the S_X's are sparse). Suppose also that there's an efficient quantum algorithm A that, for each X, samples—but does not necessarily QSample—the distribution $\mathcal{D}_X$ over y conditional on X. Then there's also an efficient quantum algorithm Q that QSamples $\mathcal{D}_X$: that is, maps $|X\rangle|0\cdots 0\rangle$ to $|X\rangle|\phi_X\rangle$ for each X (up to $\frac{1}{n^{O(1)}}$ error in trace distance).

Proof. As in Proposition 55, the algorithm Q first runs A sequentially k times, for some sufficiently large $k = n^{O(1)}$. It thereby produces the state $|\zeta_X\rangle^{\otimes k}$, where

$$|\zeta_X\rangle = \sum_{y \in S_X} \sqrt{\Pr[y|X]}\,|y\rangle\,|g_{X,y}\rangle$$

is a superposition over samples from $\mathcal{D}_X$, possibly entangled with garbage. Next, Q simulates a standard-basis measurement on the $|y\rangle$ registers of the $|\zeta_X\rangle$ states, in order to estimate an empirical frequency for each possible output string $y \in \{0,1\}^m$. (Of course, all but $n^{O(1)}$ strings will have an empirical frequency of 0 in the sample; for the sake of efficiency, the 0-frequency strings are not explicitly recorded.) Then, using these empirical frequencies, Q prepares the state $|\phi_X\rangle$ to $\frac{1}{n^{O(1)}}$ accuracy. The efficiency of the preparation procedure follows from the fact that $|\phi_X\rangle$ has support of size $\ell = n^{O(1)}$.[21] Meanwhile, accuracy follows by a Chernoff bound and union bound, together with the assumption that k was a sufficiently large polynomial compared to ℓ. As the final step, Q uses $A^\dagger$ to uncompute the $|\zeta_X\rangle$'s. ∎

As a small special case of Theorem 56, take $\ell = 2$ and $m = 1$. Then each $|\phi_X\rangle$ has the form $\alpha_X|1\rangle + \beta_X|2\rangle$, so the algorithm A could be seen as a PromiseBQP decision procedure, which accepts an input X with probability $|\beta_X|^2$ (not necessarily bounded away from $\frac{1}{2}$). We have shown that a probabilistic oracle for this decision procedure can be safely implemented up to $\frac{1}{p(n)}$ accuracy in polynomial time, for any polynomial p. A reasonable interpretation of this[22] is that $\mathsf{BQP}^{\mathsf{PromiseBQP}} = \mathsf{BQP}$, generalizing the result of Bennett et al. [11] that $\mathsf{BQP}^{\mathsf{BQP}} = \mathsf{BQP}$.

Note that, for some DP algorithms, given an input $X \in \{0,1\}^n$ we can just explicitly *calculate* a classical description of the desired output state $|\phi_X\rangle$, to $\frac{1}{\exp(n)}$ precision, deterministically and in time polynomial in n. If that description also gives rise to a small quantum circuit to prepare $|\phi_X\rangle$, then we can short-circuit the estimation procedure above, and can improve its accuracy from $\frac{1}{n^{O(1)}}$ to $\frac{1}{\exp(n)}$. As an example, suppose again that each desired output state $|\phi_X\rangle$ is a superposition over a sparse set of basis states, $S_X \subset \{0,1\}^m$ with $|S_X| = n^{O(1)}$. But now suppose that, given X, we can calculate both S_X (as a list of elements), and $\langle\phi_X|y\rangle$ for each $y \in S_X$ to $\frac{1}{\exp(n)}$ precision, in polynomial time. Then by using the Solovay-Kitaev Theorem (see [34]), we can clearly *prepare* the states $|\phi_X\rangle$—i.e., QSample—in polynomial time as well.

It is not clear how to generalize the above techniques to superpositions $|\phi_X\rangle$ over exponentially many basis states (or rather, to do so in any useful generality), even in cases where the individual amplitudes $\langle\phi_X|y\rangle$ and probabilities $|\langle\phi_X|y\rangle|^2$ are computable in polynomial time.

7.2 Efficiency of Shadow Tomography

What does all of this mean for the computational complexity of shadow tomography? In the QPMW algorithm of Section 6, recall that we needed two types of measurements: threshold mea-

[21] Since we only care about $\frac{1}{n^{O(1)}}$ accuracy, in this case we do not even need the Solovay-Kitaev Theorem (see [34]).

[22] That is, for some reasonable definition of what it means to query a PromiseBQP oracle on a superposition of inputs.

surements on all rounds, and L_σ (Hamming weight plus Laplace noise) type measurements on update rounds. Proposition 54 has shown that the L_σ measurements can be implemented in quantum polynomial time, provided the underlying POVMs $E_1, \ldots, E_m$ can be implemented in quantum polynomial time. Since a threshold measurement just consists of an L_σ measurement, followed by a binary threshold decision, followed by uncomputing of garbage, it follows that the threshold measurements can be implemented in quantum polynomial time as well, again assuming efficient procedures for the E_i's.

Unfortunately, this doesn't mean that QPMW runs in polynomial time overall. The first issue is just the sheer number of measurements m. Since QPMW needs one round per measurement, if m is exponentially large then QPMW will of course need exponential time.

The second issue is the need to maintain, and to do computations on, a classical description of the current hypothesis state σ_t, in the online learning procedure [7] that QPMW uses as a subroutine. If σ_t is stored explicitly, as a $d \times d$ Hermitian matrix, then this takes d^2 space, which is prohibitive if d is exponentially large. However, even if σ_t is stored only implicitly, say by a list of constraints that it satisfies, estimating expectation values $\mathrm{Tr}\left(E_i \sigma_t\right)$ will still take $d^{\Theta(1)}$ time in general.

In summary, if we ignore various low-order contributions, then the running time of QPMW is roughly $O\left(mL\right) + d^{O(1)}$, where L is an upper bound on the time needed to implement a single measurement E_i. By comparison, Aaronson's previous shadow tomography procedure [6] used roughly $O\left(mL\right) + d^{O(\log \log d)}$ time. Thus, QPMW improves the dependence on d from quasipolynomial to polynomial.

There is also later work by Brandão et al. [14], which connects shadow tomography to semidefinite programming and Gibbs states. Brandão et al. gave a shadow tomography procedure with the same sample complexity as Aaronson's, and running time $O\left(\sqrt{m}L\right) + d^{O(1)}$. Here the improvement from m to $\sqrt{m}$ came from, in essence, repeatedly doing Grover search over $E_1, \ldots, E_m$ to find an informative E_i. Thus, if we compare to Brandão et al., QPMW matches the improvement from $d^{O(\log \log d)}$ to $d^{O(1)}$, but not the improvement from m to $\sqrt{m}$. However, this is to be expected: unlike Aaronson's or Brandão et al.'s, our new shadow tomography procedure is *online*, which necessitates taking time linear in the number of measurements.

It's natural to wonder: is there some inherent barrier ruling out a shadow tomography procedure that runs in $(\log d)^{O(1)}$ time, avoiding the polynomial dependence on Hilbert space dimension d? We now show that there *is* such a barrier—at least if we insist that the shadow tomography procedure be online, or alternatively, that it be gentle. Our proof will use recent cryptographic lower bounds for differential privacy and for answering adaptively chosen queries, as well as our result that gentleness implies DP.

Hardness for gentle (even offline) shadow tomography. We use a result of Ullman [43], which shows that under plausible cryptographic assumptions, computing differentially private answers to more than $\widetilde{\Theta}\left(n^2\right)$ queries (where n is the database size) requires time $d^{\Omega(1)}$. This hardness result extends to quantum algorithms, under plausible cryptographic assumptions about their power. Moreover, the result constructs a single distribution $\mathcal{D}$ over $[d]$, such that it's hard for DP algorithms to compute accurate answers on databases that are drawn i.i.d. from $\mathcal{D}$. Using our result that gentleness implies DP, we derive a similar hardness result for *gentle* shadow tomography.

Theorem 57 (Ullman [43], quantum variant) *Suppose there exists a symmetric-key encryption scheme that, for keys of length κ, is semantically secure against $2^{\Omega(\kappa)}$-time quantum adver-*

saries. Then there is no quantum algorithm $\mathcal{A}$, running in time $d^{o(1)} \cdot poly(m)$, that receives as input a database X comprised of n items from $[d]$, and a set of $m = \widetilde{\Theta}\left(n^2\right)$ queries $E_1, \ldots, E_m$, such that:

(1) $\mathcal{A}$ is $(1, \frac{1}{10n})$-DP.

(2) For any distribution $\mathcal{D}$ over $[d]$, if X's entries are drawn i.i.d. from $\mathcal{D}$, then with all but a small constant probability over $\mathcal{A}$'s coins and the choice of X, for every $j \in [m]$, the j^{th} answer a_j computed by $\mathcal{A}$ satisfies:

$$\left| a_j - \frac{\sum_{i \in [n]} E_j(X_i)}{n} \right| < \frac{1}{2}.$$

Moreover, the queries $E_1, \ldots, E_m$ are each computable in $poly(n, \log d)$ time.

Using the fact that gentleness implies differential privacy (Theorem 5), we conclude that gentle shadow tomography is hard.

Corollary 58 *Suppose there exists a symmetric-key encryption scheme that, for keys of length κ, is semantically secure against $2^{\Omega(\kappa)}$-time quantum adversaries. Then there is no quantum shadow tomography procedure that is gentle on product states and runs in $d^{o(1)} \cdot poly\,(m)$ time. Moreover, this holds even for the classical special case of shadow tomography.*

Corollary 58 applies even to the offline setting, and to algorithms that are accurate only in the identical-state setting where the algorithm's input is a state of the form $\rho^{\otimes n}$. Moreover, it applies even for classical data and classical queries. We note that Theorem 57 and Corollary 58 extend to milder cryptographic assumptions, with a milder conclusion on the possible running time for gentle shadow tomography. Essentially, symmetric key encryption that is hard for time-$T(\kappa)$ quantum algorithms translates into hardness of differentially private data analysis for quantum algorithms that run in time $O\left(T(\kappa)^{\tau}\right)$, for a fixed constant $\tau > 0$. Similarly to Corollary 58, the existence of such encryption schemes rules out gentle shadow tomography in time $T(\log d)^{o(1)} \cdot poly\,(m, n)$.

Finally, we remark that Theorem 57 (and Corollary 58) do not rule out efficient gentle algorithms that are tailored to fixed classes of queries—even for exponentially large fixed classes.[23] Until recently, known DP hardness results for fixed query families, such as [23, 13, 33], relied on assumptions for which we have no quantum-secure candidate instantiation, such as bilinear maps or indistinguishability obfuscation. A recent result of Kowalczyk et al. [32] presents a candidate query family based on the existence of one-way functions. These results may also extend to gentle shadow tomography.

Hardness for online (even non-gentle) shadow tomography. We use a result of Steinke and Ullman [40] (building on earlier work by Hardt and Ullman [28]), showing that under plausible cryptographic assumptions, given n i.i.d. samples from a distribution $\mathcal{D}$ over $[d]$, it is computationally hard to answer more than $O(n^2)$ adaptively-chosen queries accurately. Under appropriate assumptions, this result extends to quantum algorithms, and shows hardness for time $d^{\Omega(1)}$:

[23] Theorem 57 does not apply because, for the specific queries used to instantiate the lower bound, the time needed to compute the queries grows with the database size. In particular, Theorem 57 does not rule out an efficient DP algorithm for answering all queries that can be computed by $poly(d)$-size circuits. More generally, for any fixed query family, it does not rule out the possibility of obtaining an efficient algorithm that is accurate so long as the database is large enough, and in particular larger than the representation of queries in the family.

Theorem 59 (Steinke and Ullman [40], quantum variant) *Suppose there exists a symmetric-key encryption scheme that, for keys of length κ, is semantically secure against $2^{\Omega(\kappa)}$-time quantum adversaries. Then there is no quantum algorithm, running in $d^{o(1)} \cdot m^{O(1)}$ time, that takes as input n independent samples from a distribution $\mathcal{D}$ over $[d]$, as well as $m = O\left(n^2\right)$ efficiently computable counting queries $E_1, \ldots, E_m$ that are chosen adversarially and adaptively, and correctly estimates $\mathrm{E}_{x \sim \mathcal{D}}\left[E_i\left(x\right)\right]$ to within a fixed constant error for each $i \in [m]$ in an online manner.*

Theorem 59 has the following as an immediate corollary.

Corollary 60 *Suppose there exists a symmetric-key encryption scheme that, for keys of length κ, is semantically secure against $2^{\Omega(\kappa)}$-time quantum adversaries. Then there is no shadow tomography procedure that is online and runs in $d^{o(1)} \cdot poly(m)$ time.*

Note that Corollary 60 applies even to online algorithms that are not gentle, and that work only in the "identical-state setting" (i.e., when the algorithm's input has the form $\rho^{\otimes n}$). Moreover, it applies even for the classical special case of shadow tomography. Finally, we note that just like Corollary 58, Corollary 60 extends to milder cryptographic assumptions, albeit with milder conclusions for the complexity of gentle shadow tomography.

7.3 Quantum Complexity Implication

We now observe that gentle measurements, whether or not derived from DP algorithms, have potentially useful applications in quantum algorithms and complexity. In particular, whenever we have an efficient implementation of a gentle measurement, we can turn it into a safe and efficient way to run an associated class of estimation subroutines on superpositions of inputs, without generating unwanted garbage.

As an example, let's now prove Theorem 7 from Section 1.4. In other words, let's show that without loss of generality, a BQP machine can coherently query an oracle that takes as input a description of a quantum circuit C, and that outputs an estimate of $\Pr\left[C \text{ accepts}\right]$ to within $\pm\varepsilon$, or a superposition over such estimates, for any desired additive error $\varepsilon = \frac{1}{n^{O(1)}}$. (In the sense that, for every BQP machine that queries such an oracle, there is another BQP machine that simulates the oracle on its own.) While this might seem obvious, we would not know how to prove it without a gentle measurement procedure of some kind.

Proof of Theorem 7. Let

$$\sum_{g,C} \alpha_{g,C} \left|g, C\right\rangle$$

be a state of the BQP machine, where g is garbage that we don't care about and C is a description of a quantum circuit whose acceptance probability (say, on the $\left|0 \cdots 0\right\rangle$ state) we'd like to estimate. Then as a first step, we map the above state to

$$\sum_{g,C} \alpha_{g,C} \left|g, C\right\rangle \left(C \left|0 \cdots 0\right\rangle\right)^{\otimes \ell},$$

for some suitable $\ell = n^{O(1)}$. Next we use the efficient implementation of the Laplace noise measurement L_σ (with $\sigma \gg \sqrt{\ell}$), from Proposition 54, to map the above to some state

$$\left|\psi\right\rangle \approx \sum_{g,C} \alpha_{g,C} \left|g, C\right\rangle \left(C \left|0 \cdots 0\right\rangle\right)^{\otimes \ell} \left|p_C\right\rangle. \tag{17}$$

Here p_C is an estimate of $\Pr[C|0\cdots0\rangle \text{ accepts}]$ to within $\pm\eta$ additive error—or more precisely, a Laplace superposition over estimates, one with the property that

$$\Pr\left[|p_C - \Pr[C|0\cdots0\rangle \text{ accepts}]| > K\eta\right] \leq \frac{1}{\exp\left(\Omega\left(K\right)\right)}$$

for all K. The equality (17) is only approximate because in reality, the $|p_C\rangle$ register is slightly entangled with the $(C|0\cdots0\rangle)^{\otimes\ell}$ registers. However, recall from Corollary 6 that L_σ is α-gentle on product states for some $\alpha = O(\sqrt{\ell}/\sigma)$. Thus, the damage to the $(C|0\cdots0\rangle)^{\otimes\ell}$ registers in trace distance can be upper-bounded by α, and the equality (17) also holds up to error α. So as a final step, we can simply uncompute the $C|0\cdots0\rangle$ registers, to produce a state that is α-close in trace distance to

$$\sum_{g,C} \alpha_{g,C} |g, C\rangle |p_C\rangle .$$

If we want to ensure that the above, in turn, is α-close to a superposition such that

$$|p_C - \Pr[C|0\cdots0\rangle \text{ accepts}]| \leq \varepsilon$$

with *certainty*, where ε is our original accuracy bound, then it suffices to choose η such that $K\eta \leq \varepsilon$ for some $K = O\left(\log \frac{1}{\alpha}\right)$. Working backwards, a calculation shows that it suffices to set

$$\ell = \Theta\left(\frac{1}{\alpha^2\eta^2}\right) = \Theta\left(\frac{\log^2 \frac{1}{\alpha}}{\alpha^2\varepsilon^2}\right).$$

In turn, if our BQP machine was going to make $T = n^{O(1)}$ such queries in sequence, it would suffice to set $\alpha = \Theta\left(\frac{1}{T}\right)$ for each of them, to ensure that the final output has trace distance at most (say) $\frac{1}{10}$ from what we'd obtain using an ideal oracle for approximating $\Pr[C|0\cdots0\rangle \text{ accepts}]$. ■

Though Theorem 7 is not particularly shocking, it serves as a model for a large number of results that could now be proven, using gentle measurement procedures derived from DP algorithms. I.e., for every DP algorithm that can be implemented coherently and in polynomial time, along the lines of Proposition 54, we get another way that quantum algorithms can be safely invoked as subroutines by other quantum algorithms.

One might wonder about the difference between Theorem 7 and our results from Section 7.1. In particular, why was the Laplace noise measurement L_σ needed for Theorem 7, but *not* needed for Theorem 56? The key point is that, in Theorem 7, we wanted outputs that were explicit estimates of $\Pr[C \text{ accepts}]$. And even if two estimates $p \neq p'$ are extremely close, the states $|p\rangle$ and $|p'\rangle$ will still be orthogonal. This is what necessitated using a gentle measurement, to break the entanglement between the output and computation registers, and thereby allow safe uncomputing. In Section 7.1, by contrast, we were content with outputs that were superpositions $|\phi_X\rangle$, with our estimates of probabilities implicitly encoded in $|\phi_X\rangle$'s amplitude vector. As a result, a slight error in estimating those probabilities would yield a state $|\phi'\rangle$ such that $\langle\phi'|\phi_X\rangle \approx 1$, and gentle measurement techniques were not needed (even if the results were *useful* for efficient implementation of gentle measurements).

Here is an interesting question that we leave open. Suppose a quantum algorithm has a polynomial-time quantum subroutine C, which on each input X, generates a sample from a probability distribution $\mathcal{D}_X$ supported on a sparse set $S_X \subset \{0,1\}^m$ with $|S_X| = n^{O(1)}$. Suppose also that the output we want, on each input X, is a polynomial-size approximate *description* of $\mathcal{D}_X$:

that is, a string z_X that lists approximations to those $\Pr_{\mathcal{D}_X}[y]$ values that are far from zero, or some other representation from which $\mathcal{D}_X$ could be efficiently sampled. Is there then, necessarily, an efficient way to implement a mapping of the form

$$\sum_X \alpha_X |X\rangle \rightarrow \sum_X \alpha_X |X\rangle |z_X\rangle,$$

with no garbage?

In the special case where C is a classical randomized algorithm, we can do this by first picking a single polynomial-size random string r, and then using r as C's randomness for *every* input X in the superposition, relying on amplification and the union bound to ensure that C succeeds on every X with overwhelming probability over the choice of r. This is an instance of the well-known "Adleman's trick" [8] from complexity theory, as used for example to prove the containment $\mathsf{BPP} \subset \mathsf{P/poly}$. The use of a single r avoids any unwanted entanglement between r and the $|X\rangle$ and $|z_X\rangle$ registers.

But what about the general case, where C is a quantum algorithm? Here Adleman's trick clearly won't work, so a different idea is needed: perhaps the use of a more sophisticated DP algorithm than the Laplace algorithm used to prove Theorem 7.

8 Open Problems

This paper established a new bridge between the fields of differential privacy and quantum measurement. But we've barely begun to explore what this bridge can carry. Here are a few of our favorite open problems.

Basic Questions

(1) Can we generalize our main result, to show that ε-DP on product states implies $O(\varepsilon\sqrt{n})$-gentleness on product states for *any* quantum measurement, rather than only for product measurements? One natural first step would be to prove this for LOCC measurements. Another would be to show that ε-triviality on product states implies $O(\varepsilon)$-gentleness (or even just $O(\varepsilon\sqrt{n})$-gentleness) on product states. Note that there are two questions here: first, given a measurement M that's ε-DP on product states, can we implement M (meaning, produce the correct output probabilities on *all* states, not just product states), in a way that happens to be $O(\varepsilon\sqrt{n})$-gentle when restricted to product states? And second, can we implement some *other* measurement M' that has essentially the same output probabilities as M on product states,[24] and that's also $O(\varepsilon\sqrt{n})$-gentle on product states, but that could be arbitrarily different from M on entangled states?

(2) In this paper, we used our DP/gentleness connection, together with known results from DP, to design and analyze a new quantum measurement procedure of independent interest (namely, QPMW). Can we also go in the opposite direction, and use known results from quantum measurement theory to say anything new about classical differential privacy?

(3) Does α-gentleness imply $O(\alpha)$-DP not merely for all $\alpha \ll \frac{1}{4}$, but for all $\alpha \ll \frac{1}{2}$?

[24] If M is ε-trivial, then to get a nontrivial question here, we demand relative error on product states that's *less* than ε.

(4) In quantum differential privacy, how much can we do in the "local model," wherein n users are each individually responsible for ensuring the privacy of their respective states ρ_i, by submitting an obscured state $\widetilde{\rho}_i$ to the database? Also, how does the local model relate to the model wherein we can only perform measurements on the n states separately, for example because of experimental limitations?

Shadow Tomography

(5) What is the true sample complexity of shadow tomography? Recall that this paper's upper bound had the form $(\log m)^2 (\log d)^2 / \varepsilon^{O(1)}$, where m is the number of measurements and d is the Hilbert space dimension. By contrast, the best known lower bound is $\Omega\left(\frac{\min\left\{d^2, \log m\right\}}{\varepsilon^2}\right)$ [6]. Is any dependence on d needed? Theorem 50 showed that, if a shadow tomography procedure is also *gentle* on product states, then it needs $\widetilde{\Omega}\left(\sqrt{\log m}\,(\log d)^{1/4}\right)$ samples. Meanwhile, Theorem 53 showed that if the procedure is online, then it needs $\widetilde{\Omega}\left(\sqrt{\min\{m, \log d\}}\right)$ samples. But what if we drop these additional requirements, or relax to gentleness on states of the form $\rho^{\otimes n}$? We stress that any lower bound will need to be "inherently quantum," since classically, in the offline and non-gentle setting, an $O\left(\frac{\log m}{\varepsilon^2}\right)$ upper bound holds independent of d [6].

(6) Is it possible to do shadow tomography using incoherent measurements (i.e., measuring each copy of ρ separately)? If so, this would bring shadow tomography much closer to experimental feasibility.

Composition

(7) What can we say about the composition of quantum DP algorithms (see Appendix 13 for further discussion)? In the regime where DP implies gentleness, but where the probabilities of outcomes are too small for Lemma 17 to apply, can we compose DP algorithms in a way that preserves not only accuracy, but also a multiplicative privacy guarantee? Also, outside the regime where DP implies gentleness, is there any way to get around the counterexample of Appendix 13, and compose quantum DP algorithms in a way that preserves accuracy (to say nothing about privacy)? For example, what about "non-black-box" composition methods?

(8) Does an "advanced composition theorem" (see [22]) hold for gentleness, or at least for the particular gentle measurements that arise from our connection between gentleness and DP? In other words, if we perform α-gentle measurements k times in sequence, then can we say that with high probability over the measurement outcomes, our states have been damaged by only $O(\alpha\sqrt{k})$ in trace distance, rather than $O(\alpha k)$? If so, we could likely improve the sample complexity of our QPMW shadow tomography procedure, say from $(\log m)^2 (\log d)^2 / \varepsilon^{O(1)}$ to $(\log m)^2 (\log d) / \varepsilon^{O(1)}$.

Computational Complexity

(9) Is there any example of a polynomial-time classical randomized algorithm that is ε-DP for some $\varepsilon \ll \frac{1}{\sqrt{n}}$, but does *not* give rise to a gentle measurement on product states that can be implemented in polynomial time, because of the issue with the computational complexity of QSampling discussed in Section 7? If so, are there any "natural" examples of such DP algorithms? It would be of interest to give such examples either conditionally (say, based on a cryptographic assumption), or unconditionally in the black-box model.

(10) Can we show, under some plausible cryptographic assumption, that $d^{\Omega(1)}$ computation time is needed for shadow tomography, without the additional constraints that the procedure be online or gentle?

(11) Can we generalize Theorem 7, to give more examples of how quantum algorithms can be safely invoked as subroutines by other quantum algorithms using gentle measurement procedures? What about the problem mentioned at the end of Section 7.3?

9 Acknowledgments

We thank Lijie Chen for insightful comments, including catching an error in a previous analysis of QPMW; Thomas Steinke, Uri Stemmer, and Jon Ullman for helpful conversations about lower bounds and hardness results for differential privacy and adaptive data analysis; Andris Ambainis, Mark Bun, Dana Moshkovitz, and Fabio Sciarrino for helpful conversations; and David Mestel and the anonymous reviewers for their comments.

References

[1] S. Aaronson. Quantum lower bound for the collision problem. In *Proc. ACM STOC*, pages 635–642, 2002. quant-ph/0111102.

[2] 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.

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

[5] 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.

[6] S. Aaronson. Shadow tomography of quantum states. In *Proc. ACM STOC*, pages 325–338, 2018. arXiv:1711.01053.

[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] L. Adleman. Two theorems on random polynomial time. In *Proc. IEEE FOCS*, pages 75–83, 1978.

[9] D. Aharonov and A. Ta-Shma. Adiabatic quantum state generation and statistical zero knowledge. In *Proc. ACM STOC*, pages 20–29, 2003. quant-ph/0301023.

[10] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proc. ACM STOC*, pages 1046–1059, 2016.

[11] C. Bennett, E. Bernstein, G. Brassard, and U. Vazirani. Strengths and weaknesses of quantum computing. *SIAM J. Comput.*, 26(5):1510–1523, 1997. quant-ph/9701001.

[12] C. H. Bennett, A. W. Harrow, and S. Lloyd. Universal quantum data compression via gentle tomography. *Phys. Rev. A*, 73(032336), 2006. arXiv:quant-ph/0403078.

[13] Dan Boneh and Mark Zhandry. Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation. *Algorithmica*, 79(4):1233–1285, 2017.

[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] S. L. Braunstein, C. M. Caves, N. Linden, S. Popescu, and R. Schack. Separability of very noisy mixed states and implications for NMR quantum computing. *Phys. Rev. Lett.*, 83:1054–1057, 1999. quant-ph/9811018.

[16] M. Bun, J. Ullman, and S. P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proc. ACM STOC*, pages 1–10, 2014. arXiv:1311.3158.

[17] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884, 1969.

[18] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Generalization in adaptive data analysis and holdout reuse. In *Proc. of Neural Information Processing Systems (NIPS)*, pages 2350–2358, 2015.

[19] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *Proc. ACM STOC*, pages 117–126, 2015.

[20] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of Theory of Cryptography Conference (TCC)*, pages 265–284, 2006.

[21] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[22] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *Proc. IEEE FOCS*, pages 51–60, 2010.

[23] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N. Rothblum, and Salil P. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 381–390, 2009.

[24] 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.

[25] L. Gurvits and H. Barnum. Separable balls around the maximally mixed multipartite quantum states. *Phys. Rev. A*, 68(042312), 2003. arXiv:quant-ph/0302102.

[26] 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.

[27] M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proc. IEEE FOCS*, pages 61–70, 2010.

[28] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *Proc. IEEE FOCS*, pages 454–463, 2014.

[29] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *Proc. IEEE FOCS*, pages 454–463, 2014.

[30] L. Hardy. Quantum theory from five reasonable axioms. quant-ph/0101012, 2003.

[31] 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.

[32] Lucas Kowalczyk, Tal Malkin, Jonathan Ullman, and Daniel Wichs. Hardness of non-interactive differential privacy from one-way functions. In *Advances in Cryptology - CRYPTO 2018 - 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2018, Proceedings, Part I*, pages 437–466, 2018.

[33] Lucas Kowalczyk, Tal Malkin, Jonathan Ullman, and Mark Zhandry. Strong hardness of privacy from weak traitor tracing. In *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I*, pages 659–689, 2016.

[34] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[35] K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman. The limits of post-selection generalization. In *Proc. of Neural Information Processing Systems (NIPS)*, 2018. arXiv:1806.06100.

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

[37] R.-D. Reiss. *Approximate Distributions of Order Statistics*. Springer Series in Statistics, 1989.

[38] R. M. Rogers, A. Roth, A. D. Smith, and O. Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proc. IEEE FOCS*, pages 487–494, 2016.

[39] M. Senekane, M. Mafu, and B. Taele. Privacy-preserving quantum machine learning using differential privacy. In *Proceedings of IEEE Africon*, pages 1432–1435, 2017.

[40] T. Steinke and J. Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proc. Info. Theory and Applications (ITA)*, pages 1–41, 2016.

[41] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. *CoRR*, abs/1704.03024, 2017.

[42] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv:1709.02753, 2017.

[43] J. Ullman. Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. *SIAM J. Comput.*, 45(2):473–496, 2016.

[44] S. Vadhan. The complexity of differential privacy. In Y. Lindell, editor, *Tutorials on the Foundations of Cryptography*. Springer, 2017. privacytools.seas.harvard.edu/files/privacytools/files/complexityprivacy_1.pdf.

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

[46] L. Zhou and M. Ying. Differential privacy in quantum computation. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 249–262, 2017.

10 Appendix: DP, Gentleness, and Triviality on Separable versus Entangled States

What is the relationship between a measurement's being differentially private (or trivial, or gentle) on product states, and its having those same properties on arbitrary states?

In this appendix, we'll give examples of measurements M on n qubits that are

(1) $\frac{1}{2^{(n-1)/2}}$-trivial, $\frac{1}{2^{(n-1)/2}}$-DP, and $\frac{1}{2^{(n-1)/2}}$-gentle on all product states (and indeed, on all separable mixed states), and yet

(2) extremely far from being trivial, private, or gentle on certain entangled states.

In some sense, this will answer our question "for complexity-theoretic purposes": doing nothing whatsoever on separable states, to some fixed exponential precision, is compatible with enormous departures from DP, gentleness, and triviality on entangled states.

Nevertheless, we'll then show that there's *some* level of triviality, DP, and gentleness on product states that implies the same properties on arbitrary states—but strikingly, that this would be false in quantum mechanics over $\mathbb{R}$ rather than over $\mathbb{C}$.

10.1 Separations

Our first example separates DP on product states from DP on arbitrary states.

Proposition 61 *There exists an n-qubit measurement M that's $O\left(2^{-n/2}\right)$-trivial (and hence, $O\left(2^{-n/2}\right)$-DP) on product states, but not ε-DP for any ε on arbitrary states.*

Proof. For simplicity, let n be odd, and group the first $n-1$ qubits into $\frac{n-1}{2}$ pairs. Then the measurement M will first project each of these pairs onto the Bell pair $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$. If all $\frac{n-1}{2}$ projections succeed, then M measures the n^{th} qubit in the $\{|0\rangle, |1\rangle\}$ basis and returns the result. Otherwise M returns a uniformly random bit.

Clearly, on states of the form

$$\frac{|00\rangle + |11\rangle}{\sqrt{2}} \otimes \cdots \otimes \frac{|00\rangle + |11\rangle}{\sqrt{2}} \otimes |v\rangle,$$

this measurement is not ε-DP for any ε, since (for example) it completely leaks whether $|v\rangle = |0\rangle$ or $|v\rangle = |1\rangle$.

On the other hand, we claim that M is $O\left(2^{-n/2}\right)$-DP on product states. To see this, observe that every 2-qubit product state has at most $\frac{1}{\sqrt{2}}$ projection onto the Bell pair $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$. So when we apply M to an n-qubit product state, the $\frac{n-1}{2}$ projections all succeed with probability at most $2^{-(n-1)/2}$—and if at least one projection fails, then M's output is random. Thus, if ρ and σ are any two product states, then for all $y \in \{0,1\}$,

$$\frac{\Pr\left[M\left(\rho\right) \text{ outputs } y\right]}{\Pr\left[M\left(\sigma\right) \text{ outputs } y\right]} \leq \frac{\frac{1}{2} + 2^{-(n-1)/2}}{\frac{1}{2} - 2^{-(n-1)/2}} = 1 + O\left(2^{-n/2}\right).$$

■

As a bonus, we can adapt Proposition 61 to separate DP on product states from DP on arbitrary states, even in the special case where the measurement M is mixture-of-products.

Proposition 62 *There exists an n-qubit mixture-of-products measurement M that's $\frac{1}{\exp(n)}$-trivial (or equivalently, $\frac{1}{\exp(n)}$-DP) on product states, but is not ε-DP for any $\varepsilon < \exp\left(n\right)$ on arbitrary states.*

Proof. We simply modify the measurement M from the proof of Proposition 61, so that now M tries to use each of the $\frac{n-1}{2}$ qubit pairs to violate a Bell inequality—say, by playing the so-called *CHSH game* [17], which can be won with probability $\cos^2 \frac{\pi}{8} \approx 0.85$ using the entangled state $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$, but with at most $\frac{3}{4}$ probability using any unentangled state.

If M wins at the CHSH game, on (say) at least an 0.8 fraction of the $\frac{n-1}{2}$ qubit pairs, then M returns the result of measuring the n^{th} qubit in the $\{|0\rangle, |1\rangle\}$ basis. Otherwise, M returns a uniformly random bit.

Again, on states of the form

$$\frac{|00\rangle + |11\rangle}{\sqrt{2}} \otimes \cdots \otimes \frac{|00\rangle + |11\rangle}{\sqrt{2}} \otimes |v\rangle,$$

this measurement is not ε-DP for any $\varepsilon < \exp(n)$, since it leaks whether $|v\rangle = |0\rangle$ or $|v\rangle = |1\rangle$ with all but exponentially small probability.

But again, on product states, we claim that M is $\frac{1}{\exp(n)}$-trivial. For by a Chernoff bound, whenever M is applied to a product state, the n^{th} qubit is measured with at most $\frac{1}{\exp(n)}$ probability. ■

The measurements M from Propositions 61 and 62 don't have product form, so we can't apply Theorem 5 to them to conclude automatically that they're $\frac{1}{\exp(n)}$-gentle on product states. Nevertheless, it's not hard to verify directly that they *are* $\frac{1}{\exp(n)}$-gentle on product states, and even on separable mixed states.

By contrast, Corollary 24 says that, if M is α-gentle on all states, then M is $\ln\left(\frac{1+4\alpha}{1-4\alpha}\right)$-DP on all states. But M is *not* ε-DP on all states, for any $\varepsilon > 0$ (in the case of Proposition 61) or for any $\varepsilon < \exp(n)$ (in the case of Proposition 62). So summarizing, we obtain the following corollary of Propositions 61 and 62, which dramatically separates *gentleness* on product states from *gentleness* on all states:

Corollary 63 *There exists an n-qubit measurement M that's $\frac{1}{\exp(n)}$-gentle (and indeed, $\frac{1}{\exp(n)}$-trivial) on product states and indeed on separable mixed states, but not α-gentle for any $\alpha < \frac{1}{4.01}$ on arbitrary states. We can even take this measurement to be mixture-of-products.*

From Proposition 36, together with Lemma 31, we already get that the measurement $L_{n/2}$ is $O(1/\sqrt{n})$-gentle on product states despite not being $\frac{1}{3}$-gentle on arbitrary states. However, Corollary 63 gives an exponentially more dramatic separation between gentleness on product states and gentleness on arbitrary states.

It will follow from Corollary 68, proved in Section 10.2, that these exponential separations, between triviality, DP, and gentleness on product states and the same parameters on arbitrary states, are the largest separations possible, up to the exact value of the exponential scaling factor.

Note also that the following is an immediate consequence of convexity and of Proposition 13:

Proposition 64 *If M is ε-trivial or ε-DP on all product states, then M is also ε-trivial or ε-DP respectively on all separable mixed states.*

Beware that α-gentleness on product states does *not* automatically imply α-gentleness on separable mixed states (even though in the examples above the two happened to go together); the measurement L_σ is a counterexample.

As a final remark, one might wonder whether the counterexamples of Propositions 61 and 62 and Corollary 63 have classical probabilistic analogues. In other words, is there a separation between DP on product distributions, and DP on arbitrary distributions? Or the analogous question for triviality? We now observe that the answer is no. Indeed, this is just a special case of Proposition 64 above. Every probability distribution can be written as a convex combination of product distributions (indeed, point distributions), and DP and triviality are both closed under convex combinations.[25]

[25] Again, gentleness is the outlier, failing to be closed under convex combinations. It's not hard to show, by a classical analogue of Lemma 23, that the only classical algorithms that are gentle on arbitrary distributions $\mathcal{D}$ are close to trivial. But *every* algorithm is, or can be made, gentle on classical computational basis states.

Why is the quantum case different? Because, while DP is closed under convex combinations, it's *not* closed under superpositions. The CHSH game provides one example of this: a certain measurement has a behavior on the Bell pair $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$ that's *not* a convex combination of its behaviors on the components $|00\rangle$ and $|11\rangle$—so that the measurement can fail to be DP on the superposition, despite being DP on the components. Thus, the separation between DP on product states and DP on arbitrary states is a quantum phenomenon.

10.2 Relationships

We'll now show that, despite the separating examples in the last section, a measurement's being ε-trivial on product states for *extremely* small values of ε (say, $\varepsilon \ll \frac{1}{(2\sqrt{2})^n}$), really does imply its being nearly trivial on arbitrary states (and hence DP and gentle as well). Intriguingly, we'll also show that this depends on the fact that amplitudes in quantum mechanics can be complex rather than only real.

Our first claim is that *any measurement M that accepts every product state with the same probability p, in fact accepts every state with probability p.* We do not know whether this was known before; in any case, we cannot resist including a strikingly simple proof for completeness. Our proof uses the following result of Braunstein et al. [15]:

Theorem 65 (Braunstein et al. [15]) *In any finite-dimensional tensor product Hilbert space (on any number of registers), the separable mixed states have positive density within the set of all mixed states.*

We observe the following consequence.

Theorem 66 *Suppose a measurement M is 0-trivial (or equivalently, 0-DP or 0-gentle) on all product states. Then M is 0-trivial on all states.*

Proof. If M is 0-trivial on product states, then for each possible outcome y, there is some constant p such that, for all product states $\rho = \rho_1 \otimes \cdots \otimes \rho_n$,

$$\Pr\left[M\left(\rho\right) \text{ outputs } y\right] = p.$$

So by convexity, the above holds as well for all convex combinations of product states: i.e., separable mixed states. Now

$$\Pr\left[M\left(\rho\right) \text{ outputs } y\right] = \mathrm{Tr}\left(E\rho\right)$$

for some Hermitian operator E. By Theorem 65, this means that the linear function $f\left(\rho\right) := \mathrm{Tr}\left(E\rho\right)$ equals p on a subset of positive density. But any linear function that's constant on a subset of positive density is constant everywhere, so $\mathrm{Tr}\left(E\rho\right) = p$ for all ρ. ■

Why did this depend on amplitudes being complex numbers? In quantum mechanics over $\mathbb{R}$, the result of Braunstein et al. [15] is known to be false. Let us now show that Theorem 66 is false as well. Consider the 2-outcome measurement on 2 "rebits" (i.e., real-amplitude qubits) that accepts ρ with probability $\mathrm{Tr}\left(E\rho\right)$, where

$$E = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}.$$

One can check that, for every 2-rebit pure product state $|\psi\rangle = |\psi_A\rangle \otimes |\psi_B\rangle$, we have

$$\mathrm{Tr}\left(E|\psi\rangle\langle\psi|\right) = \langle\psi|E|\psi\rangle = \frac{1}{2},$$

and hence the same is true for every 2-rebit separable mixed state. Nevertheless, this measurement accepts the entangled rebit state $\frac{|01\rangle+|10\rangle}{\sqrt{2}}$ with certainty, and rejects $\frac{|00\rangle+|11\rangle}{\sqrt{2}}$ with certainty. This is a rare example of a quantum information phenomenon that's fundamentally different for qubits and rebits.[26]

In ordinary (complex) quantum mechanics, we can even obtain a weak *quantitative* connection between DP, gentleness, and triviality on product states and the same notions on arbitrary states, by using the following result due to Gurvits and Barnum [25].

Theorem 67 ([25]) *Let ρ be any mixed state on n registers, each d-dimensional. Then the state $(1-\delta)\frac{\mathbb{I}}{d^n} + \delta\rho$ is separable, for all $\delta \leq \frac{2^{-n/2}}{d^n}$.*

Theorem 67 has the following corollary.

Corollary 68 *Suppose the measurement M, on n registers of d dimensions each, is ε-trivial on product states, for some $\varepsilon \leq \frac{1}{2(\sqrt{2}d)^n}$. Then M is $O\left((\sqrt{2}d)^n\varepsilon\right)$-trivial on all states.*

Proof. Fix some measurement outcome y corresponding to the POVM element E. Then let $p = \mathrm{Tr}\left(E\frac{\mathbb{I}}{d^n}\right)$ be the probability that M outputs y on the maximally mixed state. Set $\delta := \frac{1}{\left(\sqrt{2}d\right)^n}$, so that $\varepsilon \leq \frac{\delta}{2}$. Let ρ be an arbitrary state, and let

$$\sigma := (1-\delta)\frac{\mathbb{I}}{d^n} + \delta\rho.$$

Then σ is separable by Theorem 67. So since M is ε-trivial on product states,

$$pe^{-\varepsilon} \leq \mathrm{Tr}\left(E\sigma\right) \leq pe^{\varepsilon}.$$

Now,

$$\begin{aligned}\mathrm{Tr}\left(E\sigma\right) &= \mathrm{Tr}\left(E\left((1-\delta)\frac{\mathbb{I}}{d^n} + \delta\rho\right)\right)\\ &= (1-\delta)\,p + \delta\,\mathrm{Tr}\left(E\rho\right).\end{aligned}$$

Solving for $\mathrm{Tr}\left(E\rho\right)$, we find that

$$\left(1 - \frac{1-e^{-\varepsilon}}{\delta}\right)p \leq \mathrm{Tr}\left(E\rho\right) \leq \left(1 + \frac{e^{\varepsilon}-1}{\delta}\right)p$$

[26] In the same spirit: in complex quantum mechanics, one can recover the POVM E if one knows $\mathrm{Tr}\left(E\rho\right)$ for all product states ρ; but in real quantum mechanics, one can't—by the same counterexample E, which the product states ρ of 2 rebits fail to distinguish from the $\mathbb{I}/2$ POVM that accepts every state with probability $1/2$. This fact is a "dual" to the well-known fact that a mixed state ρ is uniquely determined by the values of $\mathrm{Tr}\left(E\rho\right)$ on all product measurements E (i.e., Hardy's "local tomography axiom" [30] holds), in complex quantum mechanics but not in real quantum mechanics. The "duality" between the two facts can be seen by interchanging the roles of the Hermitian matrices E and ρ in the expression $\mathrm{Tr}\left(E\rho\right)$.

This implies that M is β-trivial on all states, for

$$\beta = \ln\left(\frac{1+\frac{e^{\varepsilon}-1}{\delta}}{1-\frac{1-e^{-\varepsilon}}{\delta}}\right) = O\left(\frac{\varepsilon}{\delta}\right) = O\left((\sqrt{2}d)^n\varepsilon\right).$$

Here we've used the fact that $\varepsilon \leq \frac{\delta}{2}$. ■

11 Appendix: General Neighbor Relations

Given two states ρ, σ on n registers each, we called ρ and σ *neighbors* if it's possible to reach σ from ρ, or σ from ρ, by applying some superoperator to a single register only. In the special case where $\rho = \rho_1 \otimes \cdots \otimes \rho_n$ and $\sigma = \sigma_1 \otimes \cdots \otimes \sigma_n$ are both product states, this is simply equivalent to saying that we can reach σ from ρ by changing a single ρ_i. For correlated or entangled states, by contrast, it's not obvious that we should favor this definition over various alternatives.

Thus, call ρ and σ *superoperator neighbors* if they're neighbors in the sense above. Call them *unitary neighbors* if it's possible to reach σ from ρ, or equivalently ρ from σ, by applying some unitary transformation U to a single register only. And call them *conditioned neighbors* if it's possible to reach one from the other by applying a conditioned superoperator (i.e., a normalized quantum operation) to a single register. Clearly, all unitary neighbors are also superoperator neighbors, and all superoperator neighbors are also conditioned neighbors. But for general states, the three notions are easily seen to form a strict hierarchy. For example, $\frac{1}{\sqrt{2}}\left(|0^n\rangle + |1^n\rangle\right)$ and $\frac{1}{2}\left(|0^n\rangle\langle 0^n| + |1^n\rangle\langle 1^n|\right)$ are superoperator neighbors but not unitary neighbors, while $\frac{1}{\sqrt{2}}\left(|0^n\rangle + |1^n\rangle\right)$ and $|0^n\rangle$ are conditioned neighbors but not superoperator neighbors.

Nevertheless, we now prove that, for the task of defining ε-DP, switching from superoperator neighbors to unitary neighbors would change nothing of substance, while switching to conditioned neighbors would collapse our framework to triviality.

Proposition 69 *If M is ε-DP with respect to unitary neighbors, then M is also 2ε-DP with respect to superoperator neighbors (regardless of whether we mean DP on product states or on all states).*

Proof. Let ρ and σ be superoperator neighbors, which differ only on the i^{th} register. Let ξ be the state obtained by starting from either ρ *or* σ, and then applying a Haar-random unitary transformation U to the i^{th} register (which has the effect of putting that register into the maximally mixed state, $\mathbb{I}/d$). Then averaging over the possible U's and applying convexity, we have

$$\Pr\left[M\left(\rho\right) = y\right] \leq e^{\varepsilon} \Pr\left[M\left(\xi\right) = y\right]$$

and likewise

$$\Pr\left[M\left(\sigma\right) = y\right] \geq e^{-\varepsilon} \Pr\left[M\left(\xi\right) = y\right].$$

Hence

$$\Pr\left[M\left(\rho\right) = y\right] \leq e^{2\varepsilon} \Pr\left[M\left(\sigma\right) = y\right].$$

■

Proposition 70 *If M is ε-DP on all states with respect to postselected neighbors, then M is 4ε-trivial.*

Proof. Let ρ, σ be any two mixed states on n registers each. Let $|1\rangle\langle 1| \otimes \rho'$ be the result of measuring the first register of ρ in the $|1\rangle, \ldots, |d\rangle$ basis and getting the outcome $|1\rangle$, and let $|2\rangle\langle 2| \otimes \sigma'$ be the result of measuring the first register of σ in the $|1\rangle, \ldots, |d\rangle$ basis and getting the outcome $|2\rangle$. Then by assumption, for all possible outcomes y of M,

$$\Pr\left[M\left(\rho\right) = y\right] \leq e^{\varepsilon} \Pr\left[M\left(|1\rangle\langle 1| \otimes \rho'\right) = y\right],$$
$$\Pr\left[M\left(\sigma\right) = y\right] \geq e^{-\varepsilon} \Pr\left[M\left(|2\rangle\langle 2| \otimes \sigma'\right) = y\right].$$

But the state

$$\xi := \frac{|1\rangle\langle 1| \otimes \rho' + |2\rangle\langle 2| \otimes \sigma'}{2}$$

is a postselected neighbor of both $|1\rangle\langle 1| \otimes \rho'$ and $|2\rangle\langle 2| \otimes \sigma'$, since measuring the first register of ξ in the $|1\rangle, \ldots, |d\rangle$ basis can yield either. Hence

$$\Pr\left[M\left(|1\rangle\langle 1| \otimes \rho'\right) = y\right] \leq e^{\varepsilon} \Pr\left[M\left(\xi\right) = y\right],$$
$$\Pr\left[M\left(|2\rangle\langle 2| \otimes \sigma'\right) = y\right] \geq e^{-\varepsilon} \Pr\left[M\left(\xi\right) = y\right].$$

Chaining together the inequalities now yields

$$\Pr\left[M\left(\rho\right) = y\right] \leq e^{4\varepsilon} \Pr\left[M\left(\sigma\right) = y\right].$$

■

12 Appendix: Differential Privacy Beyond Product and LOCC Measurements

In this appendix, we'll give an example of a measurement M on n qubits, which is differentially private on all states, but which is provably *not* a product measurement, or even a mixture-of-products measurement. In other words, there's no way to implement M (even approximately) by measuring each qubit in a separately chosen basis, with none of the bases depending on the outcomes of measuring previous qubits. This rules out the possibility of a "structure theorem" showing that all DP measurements can be put into the restricted form that we mainly studied in the body of this paper.

Going further, we'll also give a second measurement M' that's differentially private on all n-qubit states, but which we conjecture is not even LOCC. That is, we conjecture that there's no way to implement M' using local operations and classical communication (even allowing adaptivity), and that entangling measurements on the qubits are needed.

To construct M, we'll use the following lemma.

Lemma 71 *There is no 2-qubit mixture-of-products measurement that accepts the states $|0\rangle|0\rangle$ and $|1\rangle|+\rangle$ with certainty, and that rejects $|0\rangle|1\rangle$ and $|1\rangle|-\rangle$ with certainty.*

Proof. It suffices to show that there's no product measurement; the lemma then follows by convexity.

A product measurement can be written $\{E_i \otimes F_j\}_{i\in[k],j\in[\ell]}$, for some one-qubit POVMs $E_1 + \cdots + E_k = \mathbb{I}$ and $F_1 + \cdots + F_\ell = \mathbb{I}$.

Suppose we knew that measuring the first qubit in the $\{|0\rangle, |1\rangle\}$ basis yielded the outcome $|0\rangle$. Then we'd need to accept with certainty if the second qubit was $|0\rangle$, and reject with certainty if the second qubit was $|1\rangle$. But since the F_i's must be Hermitian and positive semidefinite, the only POVMs $\{F_1, \ldots, F_\ell\}$ on the second qubit that achieve that objective are equivalent under trivial changes (i.e., relabelings and adding "dummy" POVM elements) to

$$F_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad F_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

—in other words, simply measuring the second qubit in the $\{|0\rangle, |1\rangle\}$ basis. Likewise, if we knew that measuring the first qubit in the $\{|0\rangle, |1\rangle\}$ basis yielded the outcome $|1\rangle$, then we'd need to accept with certainty if the second qubit was $|+\rangle$, and reject with certainty if the second qubit was $|-\rangle$. The only POVMs that achieve *that* objective are equivalent under trivial changes to

$$F'_1 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad F'_2 = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

—in other words, measuring the second qubit in the $\{|+\rangle, |-\rangle\}$ basis. But since we don't get to choose $\{F_1, \ldots, F_\ell\}$ based on the outcome of measuring the first qubit, we can't achieve both objectives simultaneously.

Finally, if the first qubit was *not* measured in the $\{|0\rangle, |1\rangle\}$ basis, but in some other basis, then the situation is "even worse," since some outcome E_i of measuring the first qubit will be compatible with the first qubit having been $|0\rangle$ *or* with its having been $|1\rangle$. So even fixing E_i, we'll again need POVM elements equivalent to F_1, F_2 *and* POVM elements equivalent to F'_1, F'_2, which contradicts $F_1 + \cdots + F_\ell = \mathbb{I}$. ■

By compactness considerations, a corollary of Lemma 71 is that there must be some constant $\eta > 0$ (we have not worked out its value) such that no mixture-of-products measurement can distinguish $|0\rangle|0\rangle$ and $|1\rangle|+\rangle$ from $|0\rangle|1\rangle$ and $|1\rangle|-\rangle$ even with success probability $1 - \eta$.

Using Lemma 71, we now prove the main result.

Theorem 72 (Existence of Non-Product Quantum DP Measurements) *There exists a measurement M on n qubits that's $O\left(\frac{\log n}{n}\right)$-DP on all states, but that cannot be approximated (say, to $\frac{1}{3}$ variation distance in the distribution over measurement outcomes) by any mixture-of-products measurement.*

Proof. Set $k := C \log n$ for some constant C. Then the measurement $M = M_\sigma$ does the following:

(1) Group the n qubits into n/k blocks $B_1, \ldots, B_{n/k}$, each of size k

(2) Within each block B_i:

- Group the qubits into pairs
- Measure each pair in the basis $\{|0\rangle|0\rangle, |0\rangle|1\rangle, |1\rangle|+\rangle, |1\rangle|-\rangle\}$
- Count the number of these measurements that return either $|0\rangle|0\rangle$ or $|1\rangle|+\rangle$, and calculate the parity of this number, $b_i \in \{0, 1\}$

(3) Return the sum $\Gamma = b_1 + \cdots + b_{n/k}$, across all n/k blocks, plus Laplace noise with average magnitude σ

Our first claim is that M is $1/\sigma$-DP on all states. This is a simple consequence of Proposition 4.

Our second claim is that there exists a probability distribution $\mathcal{D}$ over n-qubit states (in fact, product states), such that given a state ρ drawn from $\mathcal{D}$, no mixture-of-products measurement can return a nontrivial estimate of Γ. This $\mathcal{D}$ is defined as follows: first set $\Gamma := 0$ or $\Gamma := n/k$, both with equal probability $1/2$. Then let ρ be a tensor product of pairs of qubits of the form $|0\rangle|0\rangle, |0\rangle|1\rangle, |1\rangle|+\rangle, |1\rangle|-\rangle$, which is chosen uniformly at random among all such tensor products that are consistent with the chosen value of Γ.

To prove the claim, it suffices to show that no mixture-of-products measurement can guess even a *single* parity b_i, by measuring the i^{th} block, with bias more than (say) $1/n^3$ over chance. For this we appeal to Lemma 71, which says that for each pair of qubits, the measurement *cannot* perfectly distinguish whether the pair is in the state $\frac{|00\rangle\langle 00|+|1+\rangle\langle 1+|}{2}$ (which flips the parity b_i), or the state $\frac{|01\rangle\langle 01|+|1-\rangle\langle 1-|}{2}$ (which has no effect on b_i). Rather, it can only distinguish these two mixed states only with some constant rate of noise $\eta > 0$. So the situation is equivalent to the following: we are trying to guess the parity $|x| \bmod 2$ of an arbitrary $k/2$-bit string x, but each bit of x can be read only noisily, and has either an η probability of appearing as 1 despite being 0 or vice versa (with the errors independent across the bits). In such a situation, regardless of the value of $\eta > 0$, it is well-known that our bias in guessing the parity of x falls off like $1/\exp(k)$. By simply setting $k := C \log n$ for some sufficiently large constant k, we can make this bias less than $1/n^3$.

Finally, we just need to choose (say) $\sigma := \frac{n}{10k}$. In that case, the measurement $M = M_\sigma$ is ε-DP for $\varepsilon = \frac{1}{\sigma} = O\left(\frac{\log n}{n}\right)$, and it returns a nontrivial estimate of Γ. By contrast, no mixture-of-products measurement returns a nontrivial estimate of Γ, or even distinguishes the case $\Gamma = 0$ from the case $\Gamma = n/k$ with bias (say) $\Omega\left(1/n^2\right)$. ■

The proof of Theorem 72 exploited the fact that, even though differential privacy is clearly associated with a lack of "sensitivity" on the measurement's part (i.e., changing a single register can't change the output by much), this is still compatible with *local subproblems* solved by the measurement being exquisitely sensitive to local changes. That's what happens with the noisy sum of parities example: each parity is maximally sensitive to local changes, even though a noisy sum of them is not.

Now suppose we want to show something stronger: namely, that there's an n-qubit measurement M that's differentially private, but that isn't even LOCC (that is, cannot be implemented using separate measurements on each qubit, even with adaptivity). We now propose a modification M' of the measurement M from the proof of Theorem 72, which we conjecture has the required property.

Set $k := C \log n$ for some constant C. Then the measurement $M' = M'_\sigma$ does the following.

(1) Group the n qubits into n/k^3 blocks $B_1, \ldots, B_{n/k^3}$, each of size k^3

(2) Within each block B_i:

- Group the qubits into k sub-blocks $S_1, \ldots, S_k$, each of size k^2
- Within each sub-block S_j:
  - Group the qubits into pairs
  - Perform the swap test on each pair (note: the swap test accepts the product state $|v\rangle|w\rangle$ with probability equal to $\frac{1+|\langle v|w\rangle|^2}{2}$)

- Call S_j "accepting" if *every* swap test accepts, or "rejecting" if at least one of them rejects

- Count the number of accepting sub-blocks, and let b_i be the parity of this number

(3) Return the sum $\Gamma = b_1 + \cdots + b_{n/k^3}$, across all n/k^3 blocks, plus Laplace noise with average magnitude σ

Just like in the proof of Theorem 72, it's easy to see that M' is $1/\sigma$-DP on all states. Our conjecture is that M' cannot be implemented, even approximately, using LOCC measurements on the qubits. The intuition is that, if we're restricted to LOCC, then at best we can simulate each swap test imperfectly: for example, using a measurement that accepts the product state $|v\rangle |w\rangle$ with probability equal to $\frac{1+|\langle v|w\rangle|^2}{3}$, rather than $\frac{1+|\langle v|w\rangle|^2}{2}$. This would imply that we can't reliably distinguish the following two cases:

(1) within a given sub-block, *every* swap test accepts with probability 1, versus

(2) within that sub-block, $10 \log n$ swap tests accept with probability $1/2$ (i.e., the two qubits are in orthogonal states), while the remaining swap tests accept with probability 1.

For the difference between these two cases will get "lost in the Gaussian noise," which is of order $\sqrt{k} \gg 10 \log n$ if the constant C was sufficiently large. By contrast, if we take an AND of "true" swap tests, then we accept with probability 1 in case (1), versus with probability $\left(\frac{1}{2}\right)^{10 \log n} = \frac{1}{n^{10}}$ in case (2).

But if we can't reliably distinguish these cases using LOCC, then certainly we can't guess the parity, across all k sub-blocks within a given block, with bias more than $1/\exp(k)$ over random. (Whereas by contrast, using "true" swap tests, we can compute the parity across the sub-blocks with success probability $1 - \frac{1}{n^{O(1)}}$, given the promise that every sub-block satisfies either (1) or (2) above.)

If so, then the end result is that, using LOCC measurements, we can't compute the sum of the parities across the n/k^3 blocks even noisily, whereas using true swap tests, we can.

13 Appendix: On Composition of Quantum DP Algorithms

One of the central properties of classical differential privacy is that it nicely composes: that is, if we run an ε_1-DP algorithm followed by an ε_2-DP algorithm on the same database, then the resulting algorithm is $(\varepsilon_1 + \varepsilon_2)$-DP. Furthermore, *advanced composition* [22] shows that, with overwhelming probability, the loss in privacy when we compose k algorithms is even slower than linear, growing only like $\sqrt{k}$.

This immediately raises a question: does quantum differential privacy similarly compose? Here we face a new difficulty, not present in classical case: namely, when we compose quantum DP algorithms, each algorithm will in general damage our state. And this might cause not only a catastrophic loss in privacy, but even a catastrophic loss in *accuracy*.

Fortunately, we can use our connection between DP and gentleness address the concern about accuracy, at least in the regime where that connection applies. For certainly *gentleness* composes. That is, if we apply an α_1-gentle measurement M_1 followed by an α_2-gentle measurement M_2, then the result will be $(\alpha_1 + \alpha_2)$-gentle, by the triangle inequality for trace distance. And by Corollary

16, this is true even if M_2 is guaranteed to be gentle only on the *original* state (for example, because it's a product state), and not necessarily on the post-measurement states that result from applying M_1. We even conjecture that an "advanced composition" property holds for gentleness (see Section 8).

Thus, suppose we want to compose product measurements $M_1, \dots, M_k$, that are each ε-DP on product states, for some $\varepsilon \ll \frac{1}{k\sqrt{n}}$. Then by Theorem 5, these measurements are each $O\left(\varepsilon\sqrt{n}\right)$-gentle. So we can compose them while preserving good accuracy.

Even here, though, there's a potential issue with privacy. The issue arises because the later M_i's are applied not to our original state ρ, but to damaged versions of the state. And particularly if this damage is additive rather than multiplicative, we have no guarantee that the later M_i's will preserve DP (with respect to the *original* state ρ) when applied to the damaged versions. Indeed, ensuring privacy would require saying at least *something* about the post-measurement states. If, for example, we implemented some M_i in a way that gratuitously "amplified" the information in (say) the first register, copying it into the other $n-1$ registers as a byproduct of the measurement procedure, then privacy need not be preserved when we apply M_{i+1}. On the other hand, it seems plausible to us that ε-DP measurements, for $\varepsilon \ll \frac{1}{\sqrt{n}}$, can always be implemented in such a way that privacy is preserved under composition.

By using Lemma 17, the following proposition confirms that quantum DP composition works at least in the special case where we're composing a small number of quantum DP algorithms that are gentle, and all of whose outputs have appreciably large probabilities on all states.

Proposition 73 (Limited Composition for Quantum DP) *Let $\mathcal{M}$ be the sequential composition of k measurements $M_1, \dots, M_k$, where M_i is ε_i-DP on product states and α_i-gentle on product states. Suppose that for all product states ρ and all possible sequences $Y = (y_1, \dots, y_k)$ of measurement outcomes, we have*

$$\Pr\left[M_1\left(\rho\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\rho\right) \text{ outputs } y_k\right] \geq p,$$

where $p \gg \alpha_1 + \cdots + \alpha_k$. Then $\mathcal{M}$ achieves a relative accuracy of $\frac{\alpha_1+\cdots+\alpha_k}{p}$, in the sense that

$$\left| \frac{\Pr\left[\mathcal{M}\left(\rho\right) \text{ outputs } Y\right]}{\Pr\left[M_1\left(\rho\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\rho\right) \text{ outputs } y_k\right]} - 1 \right| \leq \frac{\alpha_1 + \cdots + \alpha_k}{p}$$

for all product states ρ and all Y, and in addition is ε-DP on product states for

$$\begin{aligned} \varepsilon &= \varepsilon_1 + \cdots + \varepsilon_k + \ln\left(\frac{p + (\alpha_1 + \cdots + \alpha_k)}{p - (\alpha_1 + \cdots + \alpha_k)}\right) \\ &= \varepsilon_1 + \cdots + \varepsilon_k + O\left(\frac{\alpha_1 + \cdots + \alpha_k}{p}\right). \end{aligned}$$

Proof. The relative accuracy part follows immediately from the first part of Lemma 17, which tells us that

$$\left|\Pr\left[M_1\left(\rho\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\rho\right) \text{ outputs } y_k\right] - \Pr\left[\mathcal{M}\left(\rho\right) \text{ outputs } Y\right]\right| \leq \alpha_1 + \cdots + \alpha_k.$$

For the ε-DP part, for all neighboring product states ρ, σ and all $y_1, \ldots, y_k$ we have

$$\begin{aligned}
\Pr\left[\mathcal{M}\left(\rho\right) \text{ outputs } Y\right] &\leq \left(1 - \frac{\alpha_1 + \cdots + \alpha_k}{p}\right) \Pr\left[M_1\left(\rho\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\rho\right) \text{ outputs } y_k\right] \\
&\leq \frac{e^{\varepsilon_1 + \cdots + \varepsilon_k}}{1 - \frac{\alpha_1 + \cdots + \alpha_k}{p}} \Pr\left[M_1\left(\sigma\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\sigma\right) \text{ outputs } y_k\right] \\
&\leq e^{\varepsilon_1 + \cdots + \varepsilon_k} \frac{1 + \frac{\alpha_1 + \cdots + \alpha_k}{p}}{1 - \frac{\alpha_1 + \cdots + \alpha_k}{p}} \Pr\left[\mathcal{M}\left(\sigma\right) \text{ outputs } Y\right].
\end{aligned}$$

■

In Proposition 73, product states could have been replaced by any other set of states that's closed under the neighbor relation. For the special case of product states, though, we can combine Proposition 73 with part (2) of Theorem 5 to get the following corollary, which does not need the gentleness of the underlying measurements as a separate assumption.

Corollary 74 *Let $\mathcal{M}$ be the sequential composition of k product measurements $M_1, \ldots, M_k$, each on n registers. Suppose that each M_i is ε_i-DP on product states, where $\varepsilon := \varepsilon_1 + \cdots + \varepsilon_k$ is at most $\frac{1}{10\sqrt{n}}$. Suppose also that for all product states ρ, all $i \in [k]$, and all measurement outcomes y, we have*

$$\Pr\left[M_i\left(\rho\right) \text{ outputs } y\right] \geq p_i,$$

where $p := p_1 \cdots p_k$ is at least $10\varepsilon\sqrt{n}$. Then $\mathcal{M}$ achieves a relative accuracy of $O\left(\frac{\varepsilon\sqrt{n}}{p}\right)$, in the sense that

$$\left|\frac{\Pr\left[\mathcal{M}\left(\rho\right) \text{ outputs } y_1, \ldots, y_k\right]}{\Pr\left[M_1\left(\rho\right) \text{ outputs } y_1\right] \cdots \Pr\left[M_k\left(\rho\right) \text{ outputs } y_k\right]} - 1\right| = O\left(\frac{\varepsilon\sqrt{n}}{p}\right)$$

for all product states ρ and outcomes $y_1, \ldots, y_k$, and in addition is $O\left(\frac{\varepsilon\sqrt{n}}{p}\right)$-DP on product states.

In the remainder of this appendix, we will show that, when ε is large compared to $\frac{1}{\sqrt{n}}$, so that we're outside the range where DP implies gentleness, the composition of ε-DP measurements need not even preserve *accuracy*.

Recall the "randomized response" algorithm R_β from Section 5.1, which for each $i \in [n]$ independently, simply measures the i^{th} qubit in the $\{|0\rangle, |1\rangle\}$ basis, and returns the measurement outcome with probability $\frac{1}{2} + \beta$, or its complement with probability $\frac{1}{2} - \beta$. (Thus, the output of R_β is an n-bit string.) We now give our example:

Theorem 75 (Failure of Composition for Quantum DP) *There exist n-qubit measurements M_1 and M_2 that are individually ε-DP on product states for $\varepsilon = O\left(\frac{1}{n^{1/4}}\right)$, but such that no implementation of M_1 leaves us with a post-measurement state allowing an accurate result to be returned if we later run M_2 (even supposing that we don't condition on the outcome of M_1).*

Proof. Let M_1 be the randomized response algorithm R_ε, which is $O\left(\varepsilon\right)$-DP by Proposition 33. Also, let M_2 be the variant of the L_σ mechanism from before, but in the $\{|+\rangle, |-\rangle\}$ basis. In other words, M_2 returns the number of $|+\rangle$'s plus Laplace noise of mean σ. We've seen that L_σ, and hence M_2, is $\frac{1}{\sigma}$-DP.

Now suppose that in reality, each qubit is either $|+\rangle$ or $|-\rangle$. Then by a straightforward calculation, M_1 damages each qubit by $\Theta\left(\varepsilon^2\right)$ in trace distance, even if we average over both possible measurement outcomes, by decreasing the magnitudes of the off-diagonal density matrix entries by $\Theta\left(\varepsilon^2\right)$. In more detail, the effect is simply: every $|+\rangle$ qubit flips to $|-\rangle$, and every $|-\rangle$ qubit flips to $|+\rangle$, with independent probability $\Theta\left(\varepsilon^2\right)$.

So now consider what happens when we run M_2. If we have an n-bit string x, and every bit of x gets flipped with independent probability δ, then from the corrupted string x', we can estimate the Hamming weight of the original string x to within an additive error of $\Theta\left(\sqrt{n\delta\left(1-\delta\right)}\right)$. In our case, $\delta = \Theta\left(\varepsilon^2\right)$, so this additive error is $\Theta\left(\varepsilon\sqrt{n}\right)$.

But recall that M_2 needs to estimate the number $|+\rangle$ qubits to within an additive error of $\Theta\left(1/\varepsilon\right)$. If $\frac{1}{\varepsilon} \ll \varepsilon\sqrt{n}$, or equivalently $\varepsilon \gg \frac{1}{n^{1/4}}$, then this is impossible. ■

Of course, the above leaves many further questions that one could explore: for example, what happens for ε in the range between $\frac{1}{\sqrt{n}}$ and $\frac{1}{n^{1/4}}$? Also, what if we restrict our attention to quantum DP algorithms with only a few possible outcomes (thus ruling out randomized response applied to each qubit separately)? Finally, what if we allow our "composed" algorithm to do anything it likes to obtain the desired information, including violating the specified order (e.g., applying L_σ^+ before R_ε) and even more radical changes?