Quantum chi-squared tomography and mutual information testing

Steven T. Flammia* Ryan O'Donnell†

April 17, 2024

Abstract

For quantum state tomography on rank-r dimension-d states, we show that $\widetilde{O}(r^{.5}d^{1.5}/\epsilon) \leq \widetilde{O}(d^2/\epsilon)$ copies suffice for accuracy ϵ with respect to (Bures) χ^2-divergence, and $\widetilde{O}(rd/\epsilon)$ copies suffice for accuracy ϵ with respect to quantum relative entropy. The best previous bound was $\widetilde{O}(rd/\epsilon) \leq \widetilde{O}(d^2/\epsilon)$ with respect to infidelity; our results are an improvement since infidelity is bounded above by both the relative entropy and the χ^2-divergence. For algorithms that are required to use single-copy measurements, we show that $\widetilde{O}(r^{1.5}d^{1.5}/\epsilon) \leq \widetilde{O}(d^3/\epsilon)$ copies suffice for χ^2-divergence, and $\widetilde{O}(r^2d/\epsilon)$ suffice for relative entropy.[1]

Using this tomography algorithm, we show that $\widetilde{O}(d^{2.5}/\epsilon)$ copies of a $d \times d$-dimensional bipartite state suffice to test if it has quantum mutual information 0 or at least ϵ. As a corollary, we also improve the best known sample complexity for the *classical* version of mutual information testing to $\widetilde{O}(d/\epsilon)$.

*AWS Center for Quantum Computing, and IQIM, California Institute of Technology. sflammi@amazon.com

†Computer Science Department, Carnegie Mellon University. odonnell@cs.cmu.edu

[1]Independent and contemporaneous work [14] achieved the bound $\widetilde{O}(r^2d/\epsilon)$ for infidelity. Indeed, the authors of that work were the first to show that $\widetilde{O}(d^3/\epsilon)$ is possible for infidelity with single-copy measurements.

Quantum state tomography['s] perfection is of great importance to quantum computation and quantum information.

— Nielsen and Chuang [29, p.47]

1 Introduction

Quantum state tomography — learning a d-dimensional quantum state from n copies — is a ubiquitous task in quantum information science. It is the quantum analogue of the classical task of learning a d-outcome probability distribution from n samples.

In more detail, the goal is to design an algorithm that, given $\rho^{\otimes n}$ for some (generally mixed) quantum state $\rho \in \mathbb{C}^{d \times d}$, outputs (the classical description of) an estimate[2] $\widehat{\boldsymbol{\rho}}$ that is "ϵ-close" to ρ with high probability. The main challenge is to minimize the sample (copy) complexity n as a function of d and ϵ (and sometimes other parameters, such as $r = \operatorname{rank} \rho$). We will also be concerned with the practical issue of designing algorithms that make only single-copy (as opposed to collective) measurements.

An important aspect in specifying the quantum tomography task is the meaning of "ϵ-close"; i.e., what the *loss* function is for judging the algorithm's estimate. There are many natural ways for measuring the divergence of two quantum states — even more than for two classical probability distributions — and the precise measure chosen can make a great deal of difference both to the necessary sample complexity, as well as to the utility of the final estimate for future applications.

The main goal of this paper is to show a new tomography algorithm that achieves the most stringent notion of accuracy, *(Bures) χ^2-divergence*, while having essentially the same sample complexity as previously known algorithms using *infidelity* as a loss function. We then given an application, to the *quantum mutual information testing* problem, which crucially relies on our ability to achieve efficient state tomography with respect to χ^2-divergence.

1.1 Different quantum divergences, and prior work

Let us start by recalling five important notions of "distance" between two *classical* probability distributions p, q on $[d] = \{1, 2, \dots, d\}$ (see Section 2 for more details):

$$(\ell_2^2\text{-distance}) \lesssim (\text{total variation distance})^2 \lesssim \text{Hellinger-squared } (\mathrm{H}^2) \lesssim \text{KL divergence} \lesssim \chi^2\text{-divergence}. \tag{1}$$

(Here the "$\lesssim$" ignores small constant factors.) The first of these, ℓ_2^2-distance, does not have an operational interpretation, but it is by far the easiest to calculate and reason about. The remainder are the "**big four**" [50, p.26]: **total variation (TV) distance** controls the advantage in distinguishing p from q with 1 sample; **Hellinger-squared** controls the number of samples needed to distinguish p from q with high probability; **KL divergence** has several information-theoretic interpretations; and, **$\boldsymbol{\chi^2}$-divergence** plays a central role in goodness of fit (whether an unknown p is close to a known q). We remark that the first three quantities are bounded in $[0, 1]$, but KL divergence and χ^2-divergence may be unbounded.

It is extremely easy to show (see Proposition 2.14) that, given n samples from p, the empirical estimate $\widehat{\boldsymbol{p}}$ has expected ℓ_2^2-distance at most $1/n$ from p; hence $n = O(1/\epsilon)$ samples suffices for high-probability estimation with this loss function. Moreover, Cauchy–Schwarz immediately bounds TV^2 by d times ℓ_2^2, and hence $O(d/\epsilon)$ samples suffice when ϵ denotes TV^2 (and $\Omega(d/\epsilon)$ can be proven necessary). But in fact, $n = O(d/\epsilon)$ samples suffice even when ϵ denotes the most stringent distance, χ^2-divergence. This also follows from a short calculation of the expected χ^2-divergence of $\widehat{\boldsymbol{p}}$ from p when $\widehat{\boldsymbol{p}}$ is the *add-one* empirical estimator (see Proposition 2.16).

[2] Throughout this paper we use **boldface** to denote random variables.

The preceding five distances have natural generalizations for quantum states $\rho, \sigma \in \mathbb{C}^{d \times d}$. The analogous chain of inequalities to eq. (1) is not quite true, but we have instead

$$(\text{Frobenius distance})^2 \lesssim (\text{trace distance})^2 \lesssim \text{infidelity} \lesssim \text{quantum relative entropy},\ \text{Bures } \chi^2\text{-divergence}. \tag{2}$$

While both quantum relative entropy and Bures χ^2-divergence are bounded from below by the infidelity, neither bounds the other by a constant [45]. We remark that using the "measured relative entropy" rather than the "standard" (Umegaki) quantum relative entropy *does* make the full analogous chain of inequalities hold, turning the comma above into a $\lesssim$; however, the measured relative entropy is rarely used in practice.

In the quantum case, there is a very simple empirical estimation algorithm that achieves Frobenius-squared distance ϵ with $n = O(d^2/\epsilon)$ samples (see Section 3.6); this algorithm has the additional practical merit that copies of ρ are measured individually and nonadaptively, meaning it uses n POVMs of dimension d that are fixed in advance. Kueng, Rauhut, and Terstiege [27] gave another natural algorithm of this form with a refined rank-based bound:

Theorem 1.1. *([27, Thm. 2].) There is a state tomography algorithm using nonadaptive single-copy measurements achieving expected Frobenius-squared error $O(rd/n)$ on d-dimensional states of rank at most r. Hence $n = O(rd/\epsilon)$ samples suffice to get[3] Frobenius-squared accuracy ϵ.*

Again, Cauchy–Schwarz implies that trace distance-squared is bounded by r times Frobenius-squared, so one immediately concludes that $n = O(r^2 d/\epsilon)$ copies suffice for a nonadaptive single-copy measurement algorithm achieving trace distance-squared ϵ.

Allowing for *adaptive* single-copy measurement algorithms (in which the POVM used on the tth copy of ρ may be chosen based on the outcomes of the first $t-1$ measurements), it is known that for $d = 2$ (a single qubit), $n = O(1/\epsilon)$ measurements with one "round" of adaptivity suffice for estimation with infidelity ϵ. The idea for this dates back to at least [38], with a proof appearing in, e.g., [8, Eq. 4.17]. The case of higher d is discussed in [33], but no complete mathematical analysis seems to appear in the literature.

Remark 1.2. However, prior to completing our work, we were informed by the authors of [13] that they could achieve infidelity ϵ with $\widetilde{O}(d^3/\epsilon)$ single-copy measurements and logarithmically many rounds of adaptivity.

Moving to quantum tomography algorithms that allow for a general collective measurement on all n copies, it would seem that some amount of representation theory is needed to get optimal results (intuitively, because $\rho^{\otimes n}$ lies in the symmetric subspace). The following two results were shown independently and contemporaneously:

Theorem 1.3. *([31, Cor. 1.4].) There is state tomography algorithm using collective measurements achieving expected Frobenius-squared error $O(d/n)$ on d-dimensional states. Hence $n = O(d/\epsilon)$ samples suffice to get Frobenius-squared accuracy ϵ. As a corollary of Cauchy–Schwarz, $n = O(rd/\epsilon)$ samples suffice to get trace distance-squared accuracy ϵ.*

Theorem 1.4. *([21, (14)].) There is a state tomography algorithm using collective measurements on $n = O(rd/\epsilon) \cdot \log(d/\epsilon)$ copies that achieves infidelity ϵ.*

Remark 1.5. Except for the $\log(d/\epsilon)$ factor, Theorem 1.4 is stronger than the corollary in Theorem 1.3, since $(\text{trace distance})^2 \lesssim \text{infidelity}$. If one wishes to have optimal $O(1/\epsilon)$ dependence on ϵ (no log factor), the best known result is $n = O(r^2 d/\epsilon)$ using very sophisticated representation theory [32]. On the other hand, if one wishes to have optimal $O(rd)$ dependence (no log factor), prior to the present work the best result was $O(rd/\epsilon^2)$, following from Theorem 1.3 and infidelity $\lesssim (\text{trace distance})^2$.

Turning to lower bounds, Haah–Harrow–Ji–Wu–Yu [21] showed that for collective measurements, $\Omega(d^2/\epsilon)$ samples are necessary for trace distance-squared tomography in the full-rank case, and $\Omega(\frac{rd}{\epsilon \log(d/r\epsilon)})$ are necessary in the general rank-r case; Yuen [51] recently removed the log factor in case ϵ stands for infidelity. As for single-copy measurement algorithms, [21] showed (improving on [17]) that for *nonadaptive* algorithms, $\Omega(\frac{r^2 d}{\epsilon^2 \log(1/\epsilon)})$ copies are needed for infidelity-tomography, and $\Omega(d^3/\epsilon)$ copies are needed for trace distance-squared tomography in the full-rank case. This latter bound was also very recently established [13] even in the *adaptive* single-copy case.

[3] With probability at least .99, say, by Markov's inequality.

1.2 Our results

A major question left open by the preceding results is whether quantum state tomography with $\widetilde{O}(1/\epsilon)$ dependence is possible for a notion of accuracy more stringent than that of infidelity, such as quantum relative entropy or χ^2-divergence. Although efficient learning with respect to these more stringent measures is known to be possible in the classical case, we are not aware of any previous provable results along these lines in the quantum case. Indeed, these divergences seem fundamentally more difficult to handle, not being bounded in $[0,1]$, and prior works seemed to suggest that negative results might hold for them.

Prior authors have considered tomography with respect to these stronger error notions. For example, Ferrie and Blume-Kohout [17] investigated qubit tomography with respect to quantum relative entropy, and Ref. [34] uses χ^2 hypothesis testing to study tomography of (Choi states of) quantum channels. A further motivation comes from the work of Blume-Kohout and Hayden [11], who showed that the quantum relative entropy is singled out as the unique loss function for quantum tomography once certain plausible and general desiderata of an estimator are specified.

Our main motivation, which we return to in Section 4, is a property test for zero quantum mutual information. For this application, our argument *requires* us to do quantum state tomography with respect to Bures χ^2-divergence, as only then can we use the quantum "χ^2-vs.-H^2 identity tester" from Ref. [7].

For these two stronger error notions, we essentially show that *the strongest upper bounds that one could possibly hope for indeed hold*. Our main theorem is the following:

Theorem 1.6. *Suppose there exists a tomography algorithm $\mathcal{A}$ that obtains expected Frobenius-squared error at most $f(d,r)/n$ when given n copies of a quantum state $\rho \in \mathbb{C}^{d\times d}$ of rank at most r. Then it may be transformed into a tomography algorithm $\mathcal{A}'$ that, given ϵ, r, and*

$$n = \widetilde{O}\big(\sqrt{rd}\cdot f(d,r)/\epsilon\big) \text{ copies} \quad \text{(respectively, } n = \widetilde{O}(r\cdot f(d,r)/\epsilon) \text{ copies)}, \tag{3}$$

of ρ, outputs (with probability at least .99) the classical description of a state $\widehat{\boldsymbol{\rho}}$ having

$$\mathrm{D}_{\chi^2}(\rho \,\|\, \widehat{\boldsymbol{\rho}}) \leq \epsilon \text{ Bures } \chi^2\text{-divergence accuracy} \quad \text{(respectively, } \mathrm{S}(\rho \,\|\, \widehat{\boldsymbol{\rho}}) \leq \epsilon \text{ relative entropy accuracy).} \tag{4}$$

Moreover, if $\mathcal{A}$ uses single-copy measurements, then $\mathcal{A}'$ does as well, with $O(\log 1/\epsilon)$ rounds of adaptivity.

By plugging in Theorems 1.1 and 1.3, one immediately concludes:

Corollary 1.7. *There is a state tomography algorithm using collective measurements on $n = \widetilde{O}(r^{.5}d^{1.5}/\epsilon) \leq \widetilde{O}(d^2/\epsilon)$ copies that achieves χ^2-divergence accuracy ϵ.*

Corollary 1.8. *There is a state tomography algorithm using collective measurements on $n = \widetilde{O}(rd/\epsilon)$ copies that achieves relative entropy accuracy ϵ.*

Corollary 1.9. *There is a state tomography algorithm using single-copy measurements and $O(\log 1/\epsilon)$ rounds of adaptivity on $n = \widetilde{O}(r^{1.5}d^{1.5}/\epsilon) \leq \widetilde{O}(d^3/\epsilon)$ copies that achieves χ^2-divergence accuracy ϵ.*

Corollary 1.10. *There is a state tomography algorithm using single-copy measurements and $O(\log 1/\epsilon)$ rounds of adaptivity on $n = \widetilde{O}(r^2d/\epsilon) \leq \widetilde{O}(d^3/\epsilon)$ copies that achieves relative entropy accuracy ϵ.*

Note that in the collective-measurement case, Corollary 1.8 matches (up to a logarithmic factor) the $\widetilde{O}(rd/\epsilon)$ bound known previously only for infidelity-tomography, and Corollary 1.7 also matches it in the high-rank $r = \Theta(d)$ case. As for Corollaries 1.9 and 1.10, independent and contemporaneous work [14] showed a weaker version of Corollary 1.10 with infidelity accuracy in place of relative entropy.

Remark 1.11. Although one would wish to achieve $\widetilde{O}(rd/\epsilon)$ scaling for χ^2-tomography, we later discuss in Remark 3.17 why it seems hard to achieve dependence better than $\widetilde{O}(d^{1.5}/\epsilon)$ even in the pure $r = 1$ case.

Remark 1.12. In the case of $d = 2$ (a qubit), we remove all log factors and show that $n = O(1/\epsilon)$ single-copy measurements with one round of adaptivity suffice for tomography with respect to χ^2-divergence. This simple algorithm, which illustrates the very basic idea of our Theorem 1.6, is given in Section 3.1.

Remark 1.13. Although we have suppressed polylog factors (at most quadratic) with our $\widetilde{O}(\cdot)$ notation, for the case of tomography with respect to infidelity our polylog factors are actually *better* than previously known in some regimes. As an example, for collective measurements we have an infidelity algorithm with complexity $n = \widetilde{O}(\frac{rd}{\epsilon}\log^2(1/\epsilon)\log\log(1/\epsilon))$, which improves on the $\widetilde{O}(\frac{rd}{\epsilon}\log(d/\epsilon))$ bound from [21] (and the $\widetilde{O}(\frac{rd}{\epsilon^2})$ bound following from [31]) whenever ϵ is "large"; specifically, for $\epsilon \geq \exp\bigl(-\Omega(\sqrt{q}/\log q)\bigr)$ in the q-qubit ($d = 2^q$) case. See Corollary 3.21 for details.

Finally, in Section 4 we apply our χ^2-divergence tomography algorithm to the task of *testing for zero quantum mutual information*. In this problem, the tester gets access to n copies of a *bipartite* quantum state ρ on $\mathbb{C}^A \otimes \mathbb{C}^B$ where $|A| = |B| = d$. The task is to accept (with probability at least 2/3) if the mutual information $I(A:B)_\rho$ is zero (meaning $\rho = \rho_A \otimes \rho_B$ is a product state), and to reject (with probability at least 2/3) if $I(A:B)_\rho \geq \epsilon$. We show:

Theorem 1.14. *Testing for zero quantum mutual information can be done with $n = \widetilde{O}(1/\epsilon) \cdot (d^2 + rd^{1.5} + r^{.5}d^{1.75})$ samples when ρ_A, ρ_B have rank at most $r \leq d$.*

Remark 1.15. The above bound is no worse than $\widetilde{O}(d^{2.5}/\epsilon)$, and is $\widetilde{O}(d^2/\epsilon)$ whenever $r \leq \sqrt{d}$. One should also recall the total dimension of ρ is d^2.

Remark 1.16. Harrow and Montanaro [23] have considered a related "product tester" problem in the special case where the input is a pure state $|\psi\rangle$. Whenever the maximum overlap $\langle\psi|\rho|\psi\rangle$ with any product state ρ is $1-\epsilon$, the test passes with probability $1 - \Theta(\epsilon)$ using only two copies of $|\psi\rangle$. By itself however, this bound does not test quantum mutual information in the above sense, even for the rank-1 case.

Remark 1.17. An important feature of our result is its (near-)linear scaling in $1/\epsilon$. This is despite the fact that estimating mutual information to $\pm\epsilon$ accuracy requires $\Omega(1/\epsilon^2)$ samples, even for $d = 2$ and even for the classical case.

Our proof of Theorem 1.14 has two steps. First, we learn an estimate $\widehat{\boldsymbol{\rho}}_A \otimes \widehat{\boldsymbol{\rho}}_B$ of the marginals $\rho_A \otimes \rho_B$ that has small χ^2-divergence. Then second we use the "χ^2-vs.-infidelity" state certification algorithm from [7] to test whether the unknown state ρ is close to the "known" state $\widehat{\boldsymbol{\rho}}_A \otimes \widehat{\boldsymbol{\rho}}_B$. The second step requires us to relate infidelity to relative entropy (and hence mutual information); but more crucial is that in the first step, we *must* be able to do state tomography with Bures χ^2-divergence as the loss measure. Thus we have an example where χ^2-tomography is not just done for its own sake, but is necessary for a subsequent application.

Incidentally, we also show that the same two-step process works well for the problem of testing zero *classical* mutual information given samples from a probability distribution p on $[d] \times [d]$:

Theorem 1.18. *Testing for zero classical mutual information can be done with $n = O((d/\epsilon) \cdot \log(d/\epsilon))$ samples.*

This actually improves on the best known previous algorithm, due to Bhattacharyya–Gayen–Price–Vinodchandran [10], by a factor of $d \log d$.

2 Basic results on distances and divergences

Notation 2.1. If $\rho \in \mathbb{C}^{d\times d}$ is a matrix and $S \subseteq [d]$, we will write $\rho[S] \in \mathbb{C}^{S\times S}$ for the submatrix formed by the rows and columns from S. If $S = [s] = \{1, 2, \ldots, s\}$, we will write simply $\rho[s]$. We use similar notation when $p \in \mathbb{R}^d$ is a vector.

Remark 2.2. As we frequently deal with ℓ_2^2 error or Frobenius-squared error in this work, we often use the "triangle inequality" $(a-c)^2 \leq 2(a-b)^2 + 2(b-c)^2$ without additional comment.

2.1 Classical distances and divergences

Throughout this section, let $p = (p_1, \ldots, p_d)$ and $q = (q_1, \ldots, q_d)$ denote probability distributions on $[d]$. We also use the conventions $0/0 = 0$, $x/0 = \infty$ for $x > 0$, and trust the reader to interpret other such expressions appropriately (using continuity).

We now recall some distances between probability distributions.

Definition 2.3. For $f : (0, \infty) \to \mathbb{R}$ strictly convex at 1 with $f(1) = 0$, the associated *f-divergence* is

$$\mathrm{d}_f(p \parallel q) = \mathop{\mathbf{E}}_{\boldsymbol{j} \sim q}[f(p_{\boldsymbol{j}}/q_{\boldsymbol{j}})]. \tag{5}$$

Remark 2.4. All f-divergences satisfy the data processing inequality; see e.g. [50, Thm. 4.2].

Definition 2.5. For $\alpha \in [0, \infty]$, the associated *Rényi divergence* is defined by

$$\mathrm{d}_\alpha^{\mathrm{Rén}}(p \parallel q) = \frac{1}{\alpha - 1} \ln \sum_{i=1}^{d} p_i^\alpha q_i^{1-\alpha}. \tag{6}$$

We will use a few particular cases:

Definition 2.6. The *total variation distance*, a metric, is the f-divergence with $f(x) = \frac{1}{2}|x - 1|$:

$$\mathrm{d}_{\mathrm{TV}}(p, q) = \tfrac{1}{2} \sum_{i=1}^{d} |p_i - q_i|. \tag{7}$$

Definition 2.7. The *Hellinger distance* $\mathrm{d}_{\mathrm{H}}(p, q)$, a metric, is the square-root of the f-divergence with $f(x) = (\sqrt{x} - 1)^2$:

$$\mathrm{d}_{\mathrm{H}}^2(p, q) = \sum_{i=1}^{d} (\sqrt{p_i} - \sqrt{q_i})^2. \tag{8}$$

It is also essentially a Rényi divergence. More precisely, the *Bhattacharyya coefficient* between p and q is

$$\mathrm{BC}(p, q) = \sum_{i=1}^{d} \sqrt{p_i}\sqrt{q_i} = \exp\Big(-\tfrac{1}{2} \cdot \mathrm{d}_{1/2}^{\mathrm{Rén}}(p \parallel q)\Big), \tag{9}$$

and we have $\mathrm{d}_{\mathrm{H}}^2(p, q) = 2(1 - \mathrm{BC}(p, q))$. (Note the useful tensorization identity, $\mathrm{BC}(p_1 \otimes p_2, q_1 \otimes q_2) = \mathrm{BC}(p_1, q_1) \cdot \mathrm{BC}(p_2, q_2)$.)

Definition 2.8. The *KL divergence* (or *relative entropy*) is both an f-divergence (with $f(x) = x \ln x$ or $f(x) = x \ln x - (x - 1)$) and a Rényi divergence (with $\alpha = 1$):

$$\mathrm{d}_{\mathrm{KL}}(p \parallel q) = \sum_{i=1}^{d} p_i \ln(p_i/q_i). \tag{10}$$

Also, if p is a "bipartite" probability distribution on finite outcome set $A \times B$, and p_A, p_B denote its marginals, we may define the *mutual information*

$$I(A : B)_p = \mathrm{d}_{\mathrm{KL}}(p \parallel p_A \times p_B). \tag{11}$$

Definition 2.9. The *χ^2-divergence* is the f-divergence with $f(x) = (x - 1)^2$:

$$\mathrm{d}_{\chi^2}(p \parallel q) = \sum_{i=1}^{d} \frac{(p_i - q_i)^2}{q_i} = \left(\sum_{i=1}^{d} \frac{p_i^2}{q_i}\right) - 1. \tag{12}$$

We will sometimes use the first formula even when p and/or q do not sum to 1.

Definition 2.10. The *max-relative entropy* (or *worst-case regret*) is defined to be

$$\mathrm{d}_\infty^{\mathrm{Rén}}(p \parallel q) = \max_{\substack{i \in [d] \\ p_i \neq 0}} \{\ln(p_i/q_i)\}. \tag{13}$$

The following chain of inequalities is well known (see, e.g., [20]):

Proposition 2.11. $\frac{1}{2}\mathrm{d}_{\mathrm{H}}^2(p,q) \le \mathrm{d}_{\mathrm{TV}}(p,q) \le \mathrm{d}_{\mathrm{H}}(p,q) \le \sqrt{\mathrm{d}_{\mathrm{KL}}(p \parallel q)} \le \sqrt{\mathrm{d}_{\chi^2}(p \parallel q)}$.

Some of the inequalities in the above can be slightly sharpened; e.g., one also has $\mathrm{d}_{\mathrm{TV}}(p,q) \le \sqrt{\frac{1}{2}\mathrm{d}_{\mathrm{KL}}(p \parallel q)}$, usually called *Pinsker's inequality*. Perhaps less well known is the following "reverse" form of Pinsker's inequality:

$$\mathrm{d}_{\mathrm{KL}}(p \parallel q) \le O\big(\mathrm{d}_\infty^{\mathrm{Rén}}(p \parallel q)\big) \cdot \mathrm{d}_{\mathrm{TV}}(p,q). \tag{14}$$

Moreover, it is possible to strengthen the above by putting Hellinger-squared in place of total variation distance. These facts were proven in [39]; for the convenience of the reader, we provide a streamlined proof of the following:

Proposition 2.12. *For p, q probability distributions on $[d]$ we have*

$$\mathrm{d}_{\mathrm{KL}}(p \parallel q) \le (2 + \mathrm{d}_\infty^{\mathrm{Rén}}(p \parallel q)) \cdot \mathrm{d}_{\mathrm{H}}^2(p,q). \tag{15}$$

Proof. Let us write $r_i = p_i/q_i$. Defining

$$f(r) = r\ln r - (r-1), \qquad g(r) = (\sqrt{r}-1)^2, \tag{16}$$

the elementary Lemma 2.13 proven below shows that

$$\forall r \ge 0, \qquad f(r) \le h(r)g(r), \qquad \text{where } h(r) = 2 + \max\{\ln r, 0\}. \tag{17}$$

It follows that

$$\mathrm{d}_{\mathrm{KL}}(p \parallel q) = \mathop{\mathbf{E}}_{\boldsymbol{i}\sim q}[f(r_{\boldsymbol{i}})] \le \mathop{\mathbf{E}}_{\boldsymbol{i}\sim q}[h(r_{\boldsymbol{i}})g(r_{\boldsymbol{i}})] \le \max_{i\in[d]}\{h(r_i)\} \cdot \mathop{\mathbf{E}}_{\boldsymbol{i}\sim q}[g(r_{\boldsymbol{i}})] = (2 + \mathrm{d}_\infty^{\mathrm{Rén}}(p \parallel q)) \cdot \mathrm{d}_{\mathrm{H}}^2(p,q). \tag{18}$$

□

Lemma 2.13. *Inequality (17) holds.*

Proof. Consider $a(r) \coloneqq h(r)g(r) - f(r)$. This function is continuous and piecewise differentiable on $r \ge 0$ with an exceptional point at $r = 1$. We will first show that $a(r)$ is nonnegative and increasing on $r \ge 1$. Clearly $a(1) = 0$, so we only need to show that $a'(r) \ge 0$. For $r \ge 1$, by the integral definition of the logarithm and the Cauchy–Schwarz inequality we have

$$\ln r = \int_1^r x^{-1}\mathrm{d}x \le \left(\int_1^r \mathrm{d}x\right)^{1/2}\left(\int_1^r x^{-2}\mathrm{d}x\right)^{1/2} = \sqrt{r} - \frac{1}{\sqrt{r}}\,. \tag{19}$$

Calculating the derivative of $a(r)$ and using the above inequality for the logarithm, we have that

$$a'(r) = 3 - \frac{4}{\sqrt{r}} + \frac{1}{r} - \frac{\ln r}{\sqrt{r}} \ge 3 - \frac{4}{\sqrt{r}} + \frac{1}{r} - \frac{\sqrt{r} - \frac{1}{\sqrt{r}}}{\sqrt{r}} = \frac{2\left(\sqrt{r}-1\right)^2}{r} \ge 0. \tag{20}$$

For the case $0 \le r \le 1$, we change variables to $s = 1/r$ and define $b(s) \coloneqq h(1/s)g(1/s) - f(1/s)$ for $s \ge 1$. We have $b(1) = 0$, and using again the logarithm inequality we find

$$b'(s) = \frac{2\sqrt{s} - \ln s - 2}{s^2} \ge \frac{\sqrt{s} + \frac{1}{\sqrt{s}} - 2}{s^2} = \frac{\left(\sqrt{s}-1\right)^2}{s^{5/2}} \ge 0. \tag{21}$$

□

We remark that Inequality (17) can be strengthened to $h(r) = 2 + \ln\big((2+r)/3\big)$, but as this does not change the scaling of any of our results, we will not use this stronger inequality or present our (annoyingly complicated) proof.

Finally, we mention the ℓ_2^2-distance between probability distributions, $\|p-q\|_2^2 = \sum_i (p_i - q_i)^2$. Though it does not have an operational meaning, the simplicity of computing it makes it a useful tool when analyzing other distances. For example, the χ^2-divergence is a kind of "weighted" version of ℓ_2^2-distance, in which the error term $(p_i - q_i)^2$ is weighted by $1/q_i$. We record here basic facts about estimation with respect to these distance measures.

Proposition 2.14. *Let p be a distribution on $[d]$, and suppose $\boldsymbol{q}$ is the empirical estimator formed from m samples. Then for any $S \subseteq [d]$ we have*

$$\mathbf{E}\big[\|p[S] - \boldsymbol{q}[S]\|_2^2\big] \le \|p[S]\|_1/m. \tag{22}$$

In particular, $\mathbf{E}\big[\|p - \boldsymbol{q}\|_2^2\big] \le 1/m$.

Proof. Note that $\boldsymbol{q}_i$ is distributed as Binomial$(m, p_i)/m$, hence $\mathbf{E}[(p_i - \boldsymbol{q}_i)^2] = \mathbf{Var}[\boldsymbol{q}_i] = p_i(1-p_i)/m \le p_i/m$. Summing over $i \in S$ completes the proof. □

We will also need the following variant with high-probability guarantees:

Proposition 2.15. *In the setting of Proposition 2.14, we have the following guarantees, where we introduce the notation $m_\delta = m/(c\ln(1/\delta))$ for c some large universal constant:*

(a) $\|p - \boldsymbol{q}\|_2^2 \le 1/m_\delta$ except with probability at most δ;

(b) if $\|p[S]\|_1 \ge 1/m_\delta$ then $\|\boldsymbol{q}[S]\|_1$ is within a 1.01-factor of $\|p[S]\|_1$ except with probability at most δ;

(c) if $\|p[S]\|_1 \le 1/m_\delta$ then $\|\boldsymbol{q}[S]\|_1 \le 1.01/m_\delta$ except with probability at most δ.

Proof. Items (b) and (c) follow from standard Chernoff bounds. As for Item (a), it follows from the known high-probability bound for empirically learning a distribution with respect to ℓ_2^2-error; see, e.g., [35]. We remark that it is important to use this latter result, as opposed to the generic "median-of-$O(\log(1/\delta))$-estimates" method; if we used the latter, it would be unclear how to simultaneously achieve Items (b) and (c) □

Proposition 2.16. *Fix a subset $S \subseteq [d]$ of cardinality s. Given m samples from an unknown distribution p on $[d]$, let $\boldsymbol{q}$ be the estimator formed by using the add-one estimator on elements from S, and the empirical estimator on the remaining elements. (Note that $\boldsymbol{q}$ is itself a probability distribution.) Then*

$$\mathbf{E}[\mathrm{d}_{\chi^2}(p[S] \parallel \boldsymbol{q}[S])] \le \frac{s}{m+s} + \left(\frac{(s-1)^2}{(m+1)(m+s)} - \frac{1}{m+s}\right)\|p[S]\|_1 \le s/m + (s/m)^2 \le 2s/m, \tag{23}$$

the last inequality assuming $m \ge s$. In case $S = [d]$, the sharpest upper bound above equals $\frac{d-1}{m+1} \le d/m$.

Moreover, still assuming $m \ge s$ and using the notation m_δ from Proposition 2.15, if $p_i \ge 1/m_{\delta/s}$ for all $i \in S$, then except with probability at most δ we have that $\boldsymbol{q}_i$ is within a 4-factor of p_i simultaneously for all $i \in S$.

Proof. For $i \in S$ we have that $\boldsymbol{q}_i$ is distributed as $\frac{\boldsymbol{B}+1}{m+s}$, where $\boldsymbol{B} \sim$ Binomial(m, p_i). It is elementary to show that the resulting contribution to $\mathrm{d}_{\chi^2}(p[S] \parallel \boldsymbol{q}[S])$, namely $\frac{(p_i - \boldsymbol{q}_i)^2}{\boldsymbol{q}_i}$, has expectation equal to

$$\frac{1}{m+s} + \left(\frac{(s-1)^2}{(m+1)(m+s)} - \frac{1}{m+s} - \frac{m+s}{m+1}(1-p_i)^{m+1}\right)p_i. \tag{24}$$

Dropping the term above involving $(1-p_i)^{m+1}$, and then summing over $i \in S$, yields Inequality (23).

As for the "moreover" statement, a Chernoff bound tells us that $\boldsymbol{B}$ is within a 2-factor of mp_i except with probability at most δ/s, using $mp_i \ge c\log(s/\delta)$. When this occurs, $\boldsymbol{q}_i$ is at least $p_i/4$ (using $m \ge s$) and at most $\frac{2mp_i+1}{m} \le 3p_i$ (using $c \ge 1$), so the proof is complete by a union bound over $i \in S$. □

2.2 Quantum distances and divergences

The analogous theory of distances and divergences between quantum states is quite rich [44, 26], as there are multiple quantum generalizations of both f-divergences and Rényi divergences. To distinguish between the quantum and classical cases, we use an upper-case D for quantum divergences and a lower-case d for classical divergences.

Throughout this section, let $\rho, \sigma \in \mathbb{C}^{d \times d}$ be (mixed) quantum states.

Definition 2.17. Given an f-divergence $\mathrm{d}_f(\cdot \parallel \cdot)$, the associated *measured (aka minimal) quantum f-divergence* [26] is

$$\mathrm{D}_f(\rho \parallel \sigma) = \sup_{\text{POVMs } (E_i)_{i=1}^m} \{\mathrm{d}_f(q_\rho \parallel q_\sigma)\}, \quad \text{where } q_\tau = (\mathrm{tr}(\tau E_1), \ldots, \mathrm{tr}(\tau E_m)). \tag{25}$$

Remark 2.18. All measured f-divergences satisfy the (quantum) data processing inequality. This fact follows from the definition and a reduction to the classical case.

Definition 2.19. For $\alpha \in [0, \infty]$, the associated *conventional quantum Rényi divergence* [36] is defined by

$$\mathrm{D}_\alpha^{\mathrm{Rén}}(\rho \parallel \sigma) = \frac{1}{\alpha - 1} \ln \mathrm{tr}\big(\rho^\alpha \sigma^{1-\alpha}\big). \tag{26}$$

Let us also describe a further relationship between classical and quantum Rényi entropies. To do so let us introduce the following notation:

Definition 2.20. Given the spectral decompositions

$$\rho = \sum_{i=1}^d p_i \left|\varphi_i\right\rangle\!\left\langle\varphi_i\right|, \qquad \sigma = \sum_{i=1}^d q_i \left|\psi_i\right\rangle\!\left\langle\psi_i\right|, \tag{27}$$

we define two probability distributions $P^{\rho\sigma}, Q^{\rho\sigma}$ on $[d] \times [d]$, as follows:

$$P_{ij}^{\rho\sigma} = |\langle\varphi_i|\psi_j\rangle|^2 p_i, \qquad Q_{ij}^{\rho\sigma} = |\langle\varphi_i|\psi_j\rangle|^2 q_j. \tag{28}$$

We now give a simple calculation that allows us to compute a quantum Rényi divergence from an associated classical probability distribution. This calculation has appeared in the literature as early as [30, Thm. 2.2]; see [6, Prop. 1] for an explicit statement. For convenience, we repeat the calculation here.

Proposition 2.21. $\mathrm{D}_\alpha^{\mathrm{Rén}}(\rho \parallel \sigma) = \mathrm{d}_\alpha^{\mathrm{Rén}}(P^{\rho\sigma} \parallel Q^{\rho\sigma})$.

Proof. One calculates:

$$\begin{aligned}
\mathrm{D}_\alpha^{\mathrm{Rén}}(\rho \parallel \sigma) &= \frac{1}{\alpha - 1} \ln \mathrm{tr}\left(\left(\sum_{i=1}^d p_i^\alpha \left|\varphi_i\right\rangle\!\left\langle\varphi_i\right|\right)\left(\sum_{j=1}^d q_j^{1-\alpha} \left|\psi_j\right\rangle\!\left\langle\psi_j\right|\right)\right) \\
&= \frac{1}{\alpha - 1} \ln \left(\sum_{i,j=1}^d p^\alpha |\langle\varphi_i|\psi_j\rangle|^2 q^{1-\alpha}\right) = \frac{1}{\alpha - 1} \ln \sum_{i,j=1}^d (P_{ij}^{\rho\sigma})^\alpha (Q_{ij}^{\rho\sigma})^{1-\alpha} = \mathrm{d}_\alpha^{\mathrm{Rén}}(P^{\rho\sigma} \parallel Q^{\rho\sigma}). \qquad \square
\end{aligned}$$

We now define some particular quantum distances/divergences:

Definition 2.22. The *trace distance*, a metric, is the measured f-divergence associated to total variation distance [25]:

$$\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) = \tfrac{1}{2}\|\rho - \sigma\|_1. \tag{29}$$

Definition 2.23. The *Bures distance* $\mathrm{D}_{\mathrm{B}}(\rho, \sigma)$, a metric, is the square-root of the measured f-divergence associated to Hellinger-squared [19]. It has the formula

$$\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma) = 2(1 - \mathrm{F}(\rho, \sigma)), \tag{30}$$

where $\mathrm{F}(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1$ is the *fidelity* between ρ and σ (in the "square root" convention). The *infidelity* between ρ and σ is simply $1 - \mathrm{F}(\rho, \sigma) = \frac{1}{2}\mathrm{D}_{\mathrm{B}}^2(\rho, \sigma)$.

There is a close analogy between the quantum fidelity and the classical Bhattacharrya coefficient, and indeed the analogue of Equation (9) holds if one uses the "sandwiched Rényi entropy" [28, 48]. Using instead the conventional Rényi entropy yields a slightly different notion:

Definition 2.24. The *quantum Hellinger affinity* is defined by

$$\mathrm{A}(\rho,\sigma) = \mathrm{tr}\big(\sqrt{\rho}\sqrt{\sigma}\big) = \exp\Big(-\tfrac{1}{2}\cdot \mathrm{D}^{\mathrm{Rén}}_{1/2}(\rho \parallel \sigma)\Big), \tag{31}$$

and the *quantum Hellinger distance* $\mathrm{D}_{\mathrm{H}}(\rho,\sigma)$, a metric, is defined by

$$\mathrm{D}^2_{\mathrm{H}}(\rho,\sigma) = 2(1-\mathrm{A}(\rho,\sigma)) = \mathrm{tr}\big((\sqrt{\rho}-\sqrt{\sigma})^2\big) = \|\sqrt{\rho}-\sqrt{\sigma}\|^2_{\mathrm{F}} = \mathrm{D}^2_{\mathrm{H}}(\rho,\sigma) = \mathrm{d}^2_{\mathrm{H}}(P^{\rho\sigma}, Q^{\rho\sigma}), \tag{32}$$

the last equality using Proposition 2.21. (Note also the useful tensorization identity, $\mathrm{A}(\rho_1\otimes\rho_2,\sigma_1\otimes\sigma_2) = \mathrm{A}(\rho_1,\sigma_1)\cdot\mathrm{A}(\rho_2,\sigma_2)$.)

Fortunately, the preceding two distances differ by only a small constant factor:

Fact 2.25. $\mathrm{D}^2_{\mathrm{B}}(\rho,\sigma) \le \mathrm{D}^2_{\mathrm{H}}(\rho,\sigma) \le 2\mathrm{D}^2_{\mathrm{B}}(\rho,\sigma)$.

The left inequality in Fact 2.25 is from $\mathrm{A}(\rho,\sigma) \le \mathrm{F}(\rho,\sigma)$; the right inequality follows from [6, Eq. (32)].

Definition 2.26. The *Bures χ^2-divergence* of ρ from σ is the measured f-divergence associated to classical χ^2-divergence [12, 43]. It can also be given the following formula when $\sigma = \mathrm{diag}(q_1,\dots,q_d)$ is diagonal of full rank (and this suffices to define it for general full-rank σ, since it is unitarily invariant):

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) = \sum_{i,j=1}^{d} \frac{2}{q_i+q_j}|\tau_{ij}|^2, \quad \text{where } \tau = \rho-\sigma. \tag{33}$$

We will use this formula even when $q_1,\dots,q_d \ge 0$ do not sum to 1.

Similar to the connection between ℓ_2^2-distance and χ^2-divergence in the classical case, the Bures χ^2-divergence can be seen as a kind of "weighted" version of the Frobenius-squared distance, in which the error term $|\tau_{ij}|^2$ is weighted by $\frac{2}{q_i+q_j} = \Theta(\frac{1}{\max\{q_i,q_j\}})$.

Indeed, we will frequently consider applying Equation (33) when the q_i's form (or approximately form) a nondecreasing sequence, meaning that (we expect) $q_i \le q_j$. In this case it is reasonable to use $q_i+q_j \ge q_j$, which motivates the following simple bound:

Definition 2.27. In the notation from Definition 2.26, we define

$$\widehat{\mathrm{D}}_{\chi^2}(\rho \parallel \sigma) = \sum_{i,j=1}^{d} \frac{2}{q_{\max(i,j)}}|\tau_{ij}|^2 \ge \mathrm{D}_{\chi^2}(\rho \parallel \sigma); \tag{34}$$

and, for $L = [d']$ (for $d' \le d$) we define

$$\widehat{\mathrm{D}}^{-L}_{\chi^2}(\rho \parallel \widetilde{\boldsymbol{\rho}}) = \sum_{\substack{i,j:\\ \max(i,j)\notin L}} \frac{2}{q_{\max(i,j)}}|\tau_{ij}|^2 = \sum_{k\notin L}\frac{2}{q_k}\sum_{\substack{i,j:\\ \max(i,j)=k}}|\tau_{ij}|^2. \tag{35}$$

Note that

$$\mathrm{D}_{\chi^2}(\rho \parallel \sigma) \le \mathrm{D}_{\chi^2}(\rho[L] \parallel \sigma[L]) + \widehat{\mathrm{D}}^{-L}_{\chi^2}(\rho \parallel \sigma). \tag{36}$$

Definition 2.28. The *quantum relative entropy* [46] is defined by

$$\mathrm{S}(\rho \parallel \sigma) = \mathrm{tr}(\rho(\ln\rho - \ln\sigma)) = \mathrm{D}^{\mathrm{Rén}}_1(\rho \parallel \sigma) = \mathrm{d}_{\mathrm{KL}}(P^{\rho\sigma} \parallel Q^{\rho\sigma}), \tag{37}$$

the last equality using Proposition 2.21. Also, if ρ is a "bipartite" quantum state on $A\otimes B$, where $A\cong B\cong \mathbb{C}^d$, and if ρ_A, ρ_B denote its marginals (obtained by tracing out the B, A components, respectively), the *quantum mutual information* of ρ is defined to be

$$I(A:B)_\rho = \mathrm{S}(\rho \parallel \rho_A\otimes\rho_B). \tag{38}$$

Fact 2.29. *The conventional quantum ∞-Rényi divergence (discussed in, e.g., [3]) is, by Proposition 2.21,*

$$\mathrm{D}_\infty^{\mathrm{Rén}}(\rho \parallel \sigma) = \max_{\substack{i,j\in[d] \\ \langle\varphi_i|\psi_j\rangle\neq 0}} \{\ln(p_i/q_j)\} \leq \ln\bigl(\|\rho\|\|\sigma^{-1}\|\bigr) \leq \ln\|\sigma^{-1}\|. \tag{39}$$

Remark 2.30. This quantity is *not* the same as the "quantum max-relative entropy" defined in [16]; it *would* be if one replaced the conventional Rényi entropy with its sandwiched form.

Relating some of these divergences is the following chain of inequalities:

Proposition 2.31. $\frac{1}{2}\mathrm{D}_{\mathrm{H}}^2(\rho,\sigma) \leq \mathrm{D}_{\mathrm{tr}}(\rho,\sigma) \leq \mathrm{D}_{\mathrm{B}}(\rho,\sigma) \leq \sqrt{\mathrm{S}(\rho \parallel \sigma)},\ \sqrt{\mathrm{D}_{\chi^2}(\rho \parallel \sigma)}$.

The first inequality above is from from [6, Thm. 2]. The second follows from the classical case [19]. The third also follows from the classical case and the observation that the "measured" quantum relative entropy is at most $\mathrm{S}(\rho \parallel \sigma)$ (see, e.g. [9, App. A]). The fourth also follows from the classical case, using that Bures χ^2 is the measured form of classical χ^2 [12, 43]. As with Proposition 2.11, some of these inequalities can be sharpened slightly; for example we have the quantum Pinsker inequality $\mathrm{D}_{\mathrm{tr}}(\rho,\sigma) \leq \sqrt{\frac{1}{2}\mathrm{S}(\rho \parallel \sigma)}$.

2.3 Quantum Tomography with Quantum Relative Entropy Loss

One of our main results follows easily from the above discussion of divergences. The idea is to improve on certain "reverse quantum Pinsker" results which have been studied previously; see, e.g., [5] for a quantum generalization of the reverse-Pinsker Inequality (14). We will use the following strengthened version with quantum Hellinger-squared in place of trace distance:

Theorem 2.32. *For $\rho,\sigma \in \mathbb{C}^{d\times d}$, we have $\mathrm{S}(\rho \parallel \sigma) \leq (2 + \mathrm{D}_\infty^{\mathrm{Rén}}(\rho \parallel \sigma)) \cdot \mathrm{D}_{\mathrm{H}}^2(\rho,\sigma)$.*

Proof. This is immediate from $\mathrm{S}(\rho \parallel \sigma) = \mathrm{d}_{\mathrm{KL}}(P^{\rho\sigma} \parallel Q^{\rho\sigma})$, Proposition 2.12, and Fact 2.29. □

Despite following directly from known results (up to constant factors), the above theorem does not seem to have appeared previously in the literature. Our next result shows that this can be used to automatically upgrade any quantum tomography algorithm with an infidelity guarantee to one with a relative entropy guarantee, at the expense of only a log factor (cf. our main Theorem 1.6 upgrading Frobenius-squared-tomography to χ^2-tomography).

Notation 2.33. We write Δ_ϵ for the completely depolarizing channel, which for $0 \leq \epsilon \leq 1$ acts on $\rho \in \mathbb{C}^{d\times d}$ as $\Delta_\epsilon(\rho) = (1-\epsilon)\rho + \epsilon(\mathbb{1}/d)$ (with $\mathbb{1}$ denoting the identity matrix).

Theorem 2.34. *Let $\mathcal{A}$ be a state tomography algorithm that, given n copies of $\rho \in \mathbb{C}^{d\times d}$ and parameter ϵ, outputs an estimate $\widehat{\boldsymbol{\rho}}$ achieving infidelity $\frac{1}{2}\mathrm{D}_{\mathrm{B}}^2(\rho,\widehat{\boldsymbol{\rho}}) \leq \epsilon \leq 1/2$. Then letting $\boldsymbol{\rho}' = \Delta_{2\epsilon}(\widehat{\boldsymbol{\rho}})$, we have $\mathrm{S}(\rho \parallel \boldsymbol{\rho}') \leq 16\epsilon \cdot (2 + \ln(d/2\epsilon))$.*

Applying this theorem with the previously known result of Haah–Harrow–Ji–Wu–Yu, Theorem 1.4, we immediately conclude Corollary 1.8, that there is a state tomography algorithm with respect to quantum relative entropy that has copy complexity $n = O(rd/\epsilon) \cdot \log^2(d/\epsilon)$ (using collective measurements).

Theorem 2.34 is immediate from the following (together with the fact that Hellinger-squared is upper bounded by 4 times infidelity (Fact 2.25)):

Proposition 2.35. *Suppose $\rho,\sigma \in \mathbb{C}^{d\times d}$ are quantum states with $\mathrm{D}_{\mathrm{H}}^2(\rho,\sigma) \leq \epsilon$. Then for $\sigma' = \Delta_{\epsilon/2}(\sigma)$ we have $\mathrm{S}(\rho \parallel \sigma') \leq 4\epsilon \cdot (2 + \ln(2d/\epsilon))$.*

Proof. Since $\Delta_{\epsilon/2}(\sigma)$ has smallest eigenvalue at least $\epsilon/2d$, we have $\mathrm{D}_\infty^{\mathrm{Rén}}(\rho \parallel \sigma') \leq \ln(2d/\epsilon)$ and hence from Theorem 2.32 it suffices to show $\mathrm{D}_{\mathrm{H}}^2(\rho,\sigma') \leq 4\epsilon$. In turn, since $\mathrm{D}_{\mathrm{H}}(\cdot,\cdot)$ is a metric, by Remark 2.2 it suffices to prove $\mathrm{D}_{\mathrm{H}}^2(\sigma,\sigma') \leq \epsilon$. But using Proposition 2.31, we indeed have

$$\mathrm{D}_{\mathrm{H}}^2(\sigma,\sigma') \leq \|\sigma - \sigma'\|_1 = (\epsilon/2)\|\sigma - \mathbb{1}/d\|_1 \leq (\epsilon/2)(\|\sigma\|_1 + \|\mathbb{1}/d\|_1) = \epsilon. \tag{40}$$

□

3 Quantum state tomography

We give a guide to this section:

- Section 3.1 gives a simple χ^2-tomography algorithm for qubits; it achieves copy complexity $n = O(1/\epsilon)$ (no logs) using single-copy measurements with one round of adaptivity. It serves as a small warmup for our main algorithm.
- Section 3.2 begins the main exposition of our reduction from Frobenius-squared-tomography to χ^2-tomography. This section shows how to give several useful black-box "upgrades" to any Frobenius-squared estimator.
- In Section 3.3 we give a high-level sketch of the central estimation routine for our main theorem, which takes Frobenius-squared-tomography and turns it into a χ^2-tomography algorithm "except for very small eigenvalues".
- The most involved Section 3.4 follows; it fills in all the technical details for the preceding sketch.
- Section 3.5 shows how to take the newly-established central estimation routine and massage its output to achieve either good χ^2-accuracy (with one set of parameters) or good relative entropy accuracy (with another set of parameters). It is in this section that we establish all the theorems and corollaries from Section 1.2.
- Finally, for the convenience of the reader, Section 3.6 gives a simple Frobenius-squared-tomography algorithm using single-copy measurements with complexity $n = O(d^2/\epsilon)$.

3.1 Qubit tomography with single-copy measurements

As mentioned in Section 1.1, it has long been known that one can learn a single qubit state ρ to infidelity ϵ using *single-copy* measurements on $O(1/\epsilon)$ copies of ρ, combined with one "round" of adaptivity. In this section we give a short proof of the same result but with a stronger conclusion: ϵ accuracy with respect to Bures χ^2-divergence.

We first repeat Proposition 2.16 in the simpler context of $d = 2$, and at the same time achieving a concentration bound:

Lemma 3.1. *There is a simple classical estimation algorithm that, given $n = O(\log(1/\delta)/\epsilon)$ samples from an unknown probability distribution $p = (p_0, p_1)$ on $\{0, 1\}$, outputs an estimate $\widehat{\boldsymbol{p}}$ satisfying $\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{p}}) \leq \epsilon$ except with probability at most δ.*

Proof. As shown in Proposition 2.16, if $\widehat{\boldsymbol{q}}$ is the "add-one estimator" formed from $m \geq 4/\epsilon$ samples, then $\mathbf{E}[\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{q}})] \leq \frac{1}{m+1} \leq \epsilon/4$. By Markov's inequality, the estimator is "good", meaning $\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{q}}) \leq \epsilon$, except with probability at most $1/4$. If we now use $n = O(\log(1/\delta)/\epsilon)$ samples to produce $O(1/\delta)$ independent such estimators, a Chernoff bound tells us that, except with probability at most δ, at least a $2/3$ fraction of them are "good". If we now associate each of our estimates $\widehat{\boldsymbol{q}} = (\widehat{\boldsymbol{q}}_0, \widehat{\boldsymbol{q}}_1)$ with the point $\widehat{\boldsymbol{q}}_1$ in the interval $[0, 1]$, we see that all of the "good" points appear consecutively. (That is, "reading left-to-right", the $\widehat{\boldsymbol{q}}_1$ values consist of some "bad" points, followed by some "good" points, followed by some "bad" points.) The reason for this is that $\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{q}})$ is a monotonic function of $|p_1 - \widehat{\boldsymbol{q}}_1|$. Thus if the algorithm now selects the median $\widehat{\boldsymbol{q}}_1$ point (and its associated estimate $\widehat{\boldsymbol{q}} = (\widehat{\boldsymbol{q}}_0, \widehat{\boldsymbol{q}}_1)$), this will be among the $2/3$-fraction "good" points except with probability at most δ. □

Theorem 3.2. *There is an efficient quantum state tomography algorithm that uses $n = O(\log(1/\delta)/\epsilon)$ copies of an unknown qubit state $\rho \in \mathbb{C}^{2\times 2}$ and outputs an estimate $\widehat{\boldsymbol{\rho}}$ satisfying $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \leq \epsilon$ except with probability at most δ. Moreover, the algorithm is simple to implement in the following sense: The first $n/4$ copies of ρ are separately measured in the Pauli X basis, the next $n/4$ in the Pauli Y basis, the next $n/4$ in the Pauli Z basis, and the final $n/4$ in a fixed basis determined by the first $3n/4$ measurement outcomes.*

Proof. The first phase of the algorithm (using $3n/4$ copies) can employ any standard single-copy quantum state tomography routine; the specific one we describe in Section 3.6 has the stated Pauli format, and (using Proposition 2.15) will return a PSD matrix $\boldsymbol{\rho}'$ (not necessarily a state) satisfying

$$\|\rho - \boldsymbol{\rho}'\|_{\mathrm{F}}^2 \leq \epsilon/4 \tag{41}$$

except with probability at most $\delta/2$. Next, the algorithm employs a change of basis so as to make $\boldsymbol{\rho}'$ diagonal. It suffices to estimate ρ in this new basis. Since Frobenius-distance is unitarily invariant, in the new basis Inequality (41) implies

$$|\rho_{01}|^2 = |\rho_{10}|^2 \leq \epsilon/8. \tag{42}$$

As for the diagonal entries $p = (\rho_{00}, \rho_{11})$ of ρ, the algorithm measures its remaining $n/4$ copies of ρ in the diagonal basis and employs Lemma 3.1. For $n = O(\log(1/\delta)/\epsilon)$, this produces an estimate $\widehat{\boldsymbol{p}}$ satisfying $\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{p}}) \leq \epsilon/2$ except with probability at most $\delta/2$. The final estimate of ρ (in the new basis) will be $\widehat{\boldsymbol{\rho}} = \mathrm{diag}(\widehat{\boldsymbol{p}})$.

Except with probability at most $\delta/2 + \delta/2 = \delta$, both components of the preceding algorithm produce good estimates. Then using Equation (33) we may decompose $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}})$ into the on-diagonal contribution, which is $\mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{p}}) \leq \epsilon/2$, and the off-diagonal contribution, which is $2|\rho_{01}|^2 + 2|\rho_{10}|^2 \leq \epsilon/2$ (by Inequality (42). This completes the proof. □

3.2 Upgrading Frobenius-squared tomography algorithms

Definition 3.3. A function f mapping quantum states to numbers at least 1 will be called a *rate function*.

Definition 3.4. We say a quantum state estimation algorithm $\mathcal{A}$ has *Frobenius-squared rate* f if the following holds: Whenever $\mathcal{A}$ is given $m \in \mathbb{N}^+$ as well as $\rho^{\otimes m}$ for some quantum state $\rho \in \mathbb{C}^{d \times d}$, it outputs a matrix $\widehat{\boldsymbol{\rho}} \in \mathbb{C}^{d \times d}$ (not necessarily a state) satisfying $\mathbf{E}[\|\rho - \widehat{\boldsymbol{\rho}}\|_{\mathrm{F}}^2] \leq f(\rho)/m$.

Theorems 1.1 and 1.3 may be restated as follows:

Theorem 3.5. *There is an estimation algorithm with Frobenius-squared rate $O(d)$ on d-dimensional states.*

Theorem 3.6. *There is an estimation algorithm using single-copy measurements with Frobenius-squared rate $O(rd)$ on d-dimensional states of rank at most r.*

Finally, Proposition 3.25 gives a simple single-copy measurement algorithm that has Frobenius-squared rate $O(d^2)$ (matching matching Theorem 3.6 in the high-rank case).

We will now successively describe several black-box "upgrades" one may make to a Frobenius-squared estimation algorithm. *All of these will have the feature that they preserve the single-copy measurement property.* Our ultimate goal will be to upgrade to closeness guarantees with respect to much stronger distance measures, with minimal loss in rate. To illustrate the idea, we start with a very simple upgrade (that most natural algorithms are unlikely to need):

Proposition 3.7. *A Frobenius-squared estimation algorithm may be transformed to one that always outputs Hermitian estimates, with no loss in rate.*

Proof. Given algorithm $\mathcal{A}$, let $\mathcal{A}'$ be the algorithm that on input ρ runs $\mathcal{A}$, producing $\widehat{\boldsymbol{\rho}}$, and then outputs $\widehat{\boldsymbol{\rho}}_H \coloneqq (\widehat{\boldsymbol{\rho}} + \widehat{\boldsymbol{\rho}}^\dagger)/2$, so that $\widehat{\boldsymbol{\rho}}_A \coloneqq (\widehat{\boldsymbol{\rho}} - \widehat{\boldsymbol{\rho}}^\dagger)/2$ and $\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\rho}}_H + \widehat{\boldsymbol{\rho}}_A$. The Hermitian matrices are a real vector space, so by picking a (Hilbert–Schmidt) orthogonal basis (for example, the generalized Pauli matrices), is it easy to verify that for Hermitian ρ, we always have $\|\widehat{\boldsymbol{\rho}} - \rho\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{\rho}}_H - \rho\|_{\mathrm{F}}^2 + \|\widehat{\boldsymbol{\rho}}_A\|_{\mathrm{F}}^2 \geq \|\widehat{\boldsymbol{\rho}}_H - \rho\|_{\mathrm{F}}^2$. The claim then follows by taking expectations. □

The next upgrade is not a change in algorithm, but rather in terminology.

Definition 3.8. Say that an estimation algorithm with Frobenius-squared rate f *returns diagonal estimates* if, when run on $\rho \in \mathbb{C}^{d \times d}$, it returns a unitary $\boldsymbol{U}$ and a (real) diagonal matrix $\widehat{\boldsymbol{\rho}} = \mathrm{diag}(\boldsymbol{q})$ with $\boldsymbol{q}_1 \leq \boldsymbol{q}_2 \leq \cdots \leq \boldsymbol{q}_d$ such that $\mathbf{E}[\|\boldsymbol{U}\rho\boldsymbol{U}^\dagger - \widehat{\boldsymbol{\rho}}\|_{\mathrm{F}}^2] \leq f(\rho)$.

Given such an algorithm, we can get an Frobenius-squared estimator with rate f for ρ just by returning $\boldsymbol{U}^\dagger \widehat{\boldsymbol{\rho}} \boldsymbol{U}$. But we will prefer the interpretation that the algorithm is allowed to "revise" ρ to state $\boldsymbol{U}\rho\boldsymbol{U}^\dagger$ (with $\boldsymbol{U}$ of its choosing), and then try to estimate this new state.

Proposition 3.9. *A Frobenius-squared estimation algorithm may be transformed to one that returns diagonal estimates, with no loss in rate.*

Proof. First we transform the algorithm to output Hermitian estimates, using Proposition 3.7. Then, given output $\widehat{\boldsymbol{\rho}}$, the algorithm simply chooses a unitary $\boldsymbol{U}$ such that $\widehat{\boldsymbol{\rho}} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{q})\boldsymbol{U}^\dagger$ with $\boldsymbol{q}_1 \leq \cdots \leq \boldsymbol{q}_d$, and returns the unitary $\boldsymbol{U}$ along with diagonal estimate $\mathrm{diag}(\boldsymbol{q})$. The proof is complete because Frobenius-squared distance is unitarily invariant. □

Proposition 3.10. *With only constant-factor rate loss, a Frobenius-squared estimation algorithm may be transformed to one that outputs diagonal estimates $\widehat{\boldsymbol{\rho}} = \mathrm{diag}(\boldsymbol{q})$ that are genuine quantum states, meaning that $\boldsymbol{q}$ is a probability vector.*

Proof. First we apply Proposition 3.9, obtaining algorithm $\mathcal{A}'$ with diagonal estimates and rate f. Now our transformed algorithm, when given m copies of ρ, will start by running $\mathcal{A}'$ on the first $m/2$ copies (we may assume m is even), yielding a diagonal estimate $\boldsymbol{\rho}'$ (and, to be formal, a unitary $\boldsymbol{U}$ which should be used to conjugate the remaining copies of ρ). Say that $\|\boldsymbol{\rho}' - \rho\|_{\mathrm{F}}^2 = \boldsymbol{\eta}$, and recall that $\mathbf{E}[\boldsymbol{\eta}] \leq 2f(\rho)/m$. The next step of the algorithm is to use single-copy standard basis measurements with the remaining $m/2$ copies of ρ to make a new estimate $\boldsymbol{q}$ of the diagonal of ρ. Applying Proposition 2.14, the empirical estimator $\boldsymbol{q}$ is a genuine probability distribution, and the algorithm will finally output $\widehat{\boldsymbol{\rho}} = \mathrm{diag}(\boldsymbol{q})$. (Actually, since $\boldsymbol{q}$ might not have nondecreasing entries, we should finally "revise" by a permutation matrix.) The Frobenius-squared error of $\widehat{\boldsymbol{\rho}}$ is its off-diagonal Frobenius-squared error plus its diagonal Frobenius-squared error; the former is at most $\boldsymbol{\eta}$ and the latter is, in expectation, at most $2/m$ by Proposition 2.14. Since $\mathbf{E}[\boldsymbol{\eta}] \leq 2f(\rho)/m$, the total expected Frobenius-squared error is at most $2f(\rho)/m + 2/m = O(f(\rho))/m$, as needed. □

We will also need a high-probability version of the preceding result, with some extra properties. The reader should recall the m_δ notation from Proposition 2.15.

Proposition 3.11. *The algorithm from Proposition 3.10 may be modified so that, given $0 < \delta < 1/2$, its output satisfies each of the following statements except with probability at most δ (for any fixed $i \in [d]$):*

- $\|\rho - \widehat{\boldsymbol{\rho}}\|_{\mathrm{F}}^2 \leq f/m_\delta$;
- *if $\rho_{ii} \geq 1/m_\delta$ then $\widehat{\boldsymbol{\rho}}_{ii}$ is within a 1.01-factor of ρ_{ii};*
- *if $\rho_{ii} \leq 1/m_\delta$ then $\widehat{\boldsymbol{\rho}}_{ii} \leq 1.01/m_\delta$.*

Proof. The first statement may be obtained in a black-box way using the "median trick", which upgrades estimation-in-expectation to estimation-with-confidence-$(1-\delta)$ at the expense of only an $O(\log(1/\delta))$ sample complexity factor. This trick may be applied whenever the loss measure is a metric (as Frobenius distance is); see, e.g., [22, Prop. 2.4] for details. It is sufficient to prove this statement with the $O(\cdot)$, because we may then remove it by raising c in the m_δ notation. (Similarly, we may tolerate achieving 2δ failure probabilities, rather than δ.)

To get the other two conclusions, we need to re-estimate the diagonal of ρ, just as we did in Proposition 3.10. For this we use Proposition 2.15. As in Proposition 3.10, this re-estimation contributes some new on-diagonal Frobenius-squared distance, but only at most $1/m_\delta \leq f/m_\delta$; thus the proposition's first statement remains okay. The remaining statements follow from Proposition 2.15 by taking its "S" to be $\{i\}$. □

Now we come to a most important reduction: being able to estimate *subnormalized states*. Let us define terms, and make the simplifying assumption that rate functions for proper states only depend on dimension and rank, and that they are nondecreasing functions of these parameters. We also assume for simplicity that our subnormalized states arise just from submatrices, but they could just as well arise from any given projector Π.

Definition 3.12. We say a *subnormalized* state estimation algorithm $\mathcal{A}$ has Frobenius-squared rate $f(d, r, \tau)$ if the following holds: Whenever $\mathcal{A}$ is given a subset $S \subseteq [d]$, as well as $\rho^{\otimes m}$ for some quantum state $\rho \in \mathbb{C}^{d\times d}$ of rank at most r, it outputs an estimate $\widehat{\boldsymbol{\rho}}[S] \in \mathbb{C}^{S\times S}$ such that $\mathbf{E}[\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2] \le f(d, r, \tau)/m$, where τ denotes $\operatorname{tr}\rho[S]$.

Remark 3.13. In the above definition, we may also include the condition of "returning diagonal estimates" as in Definition 3.8, with the returned unitary $\boldsymbol{U}$ being in $\mathbb{C}^{S\times S}$. Moreover, for linguistic simplicity we will henceforth assume that "diagonal estimates" are also required to have nonnegative (diagonal) entries.

Remark 3.14. Our subnormalized state estimation algorithms will actually achieve improved rate $f(d', r', \tau)$, where $d' = |S| \le d$ and $r' = \operatorname{rank}\rho[S] \le r$, but we will not try to squeeze anything out of this, for simplicity.

Proposition 3.15. *A state estimation algorithm $\mathcal{A}$ with Frobenius-squared rate $f(d, r)$ may be transformed to a subnormalized state estimation algorithm $\mathcal{A}'$, returning diagonal estimates, and having Frobenius-squared rate $f(d, r, \tau) = O(\tau \cdot f(d, r))$.*

Proof. We first apply Proposition 3.10 so that $\mathcal{A}$ may be assumed to output diagonal, genuine quantum states. This only changes bounds by constant factors on m, to which the statement of this proposition is anyway insensitive.

Given $S \subseteq [d]$ and $\rho^{\otimes m}$, let us write $\tau = \operatorname{tr}\rho[S]$ and also introduce the quantum state $\rho_{|S} = \rho[S]/\tau$ (when $\tau > 0$). The first step of the new algorithm $\mathcal{A}'$ is to measure each copy of ρ using the two-outcome PVM $(\mathbb{1}_S, \mathbb{1}_{[d]\setminus S})$. It retains all copies that have outcome S and discards the rest. In this way, $\mathcal{A}'$ obtains $(\rho_{|S})^{\otimes \boldsymbol{m}'}$, where $\boldsymbol{m}' \sim \mathrm{Binomial}(m, \tau)$. If $\boldsymbol{m}' = 0$ then the algorithm will return the 0 matrix. Otherwise, if $\boldsymbol{m}' \neq 0$ the algorithm applies $\mathcal{A}$ to $\rho_{|S}$ and obtains an estimate $\widehat{\boldsymbol{\rho}}_{|S}$ with expected Frobenius-squared error at most $f(d', r')/\boldsymbol{m}' \le f(d, r)/\boldsymbol{m}'$, where $d' = |S|$, $r' = \operatorname{rank}\rho[S]$. The final estimate that $\mathcal{A}'$ produces for $\rho[S]$ will be $\widehat{\boldsymbol{\rho}}[S] \coloneqq (\boldsymbol{m}'/m)\widehat{\boldsymbol{\rho}}_{|S}$; indeed, we can use this expression even in the $\boldsymbol{m}' = 0$ case. We now have

$$\mathbf{E}\big[\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2\big] = \mathbf{E}\big[\|\tau\rho_{|S} - (\boldsymbol{m}'/m)\widehat{\boldsymbol{\rho}}_{|S}\|_{\mathrm{F}}^2\big] = \tau^2\,\mathbf{E}\big[\|\rho_{|S} - (\boldsymbol{m}'/(\tau m))\widehat{\boldsymbol{\rho}}_{|S}\|_{\mathrm{F}}^2\big]. \tag{43}$$

We write

$$\rho_{|S} - (\boldsymbol{m}'/(\tau m))\widehat{\boldsymbol{\rho}}_{|S} = \boldsymbol{\Delta} + \boldsymbol{R}, \quad \boldsymbol{\Delta} \coloneqq \rho_{|S} - \widehat{\boldsymbol{\rho}}_{|S}, \quad \boldsymbol{R} \coloneqq (1 - \boldsymbol{m}'/(\tau m))\widehat{\boldsymbol{\rho}}_{|S}, \tag{44}$$

and use

$$\mathbf{E}\big[\|\boldsymbol{\Delta} + \boldsymbol{R}\|_{\mathrm{F}}^2\big] \le 2\,\mathbf{E}\big[\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\big] + 2\,\mathbf{E}\big[\|\boldsymbol{R}\|_{\mathrm{F}}^2\big]. \tag{45}$$

By assumption on $\mathcal{A}$, for $m' > 0$ we have

$$\mathbf{E}\big[\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 \mid \boldsymbol{m}' = m'\big] \le f(d, r)/m' \le 2f(d, r)/(m' + 1), \tag{46}$$

and this is also true even for $m' = 0$ (recall we always assume $f \ge 1$). Using the elementary fact $\mathbf{E}\big[\frac{1}{\mathrm{Bin}(m,\tau)+1}\big] = \frac{1-(1-\tau)^{m+1}}{\tau(m+1)} \le 1/(\tau m)$, we conclude

$$\mathbf{E}\big[\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2\big] \le 2f(d', r')/(\tau m). \tag{47}$$

As for $\mathbf{E}\big[\|\boldsymbol{R}\|_{\mathrm{F}}^2\big]$, let us first observe that conditioned on any $\boldsymbol{m}' = m$ (including $m = 0$), we have $\|\widehat{\boldsymbol{\rho}}_{|S}\|_{\mathrm{F}} \le 1$ with certainty, simply because $\mathcal{A}$ always outputs a genuine quantum state. Thus

$$\mathbf{E}\big[\|\boldsymbol{R}\|_{\mathrm{F}}^2\big] \le \mathbf{E}\big[(1 - \boldsymbol{m}'/(\tau m))^2\big] = (1 - \tau)/(\tau m) \le 1/(\tau m). \tag{48}$$

Combining all of the above (and using $f \ge 1$ again), we conclude $\mathbf{E}\big[\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2\big] \le \tau^2 \cdot O(f(d, r)/(\tau m))$, as needed. □

Finally, we use Proposition 3.11 to obtain some high-probability guarantees:

Proposition 3.16. *A state estimation algorithm having Frobenius-squared rate $f(d, r)$ may be transformed (preserving the single-copy measurement property) into a subnormalized state estimation algorithm returning diagonal estimates with the following properties:*

Given parameters r, δ, and $S \subseteq [d]$, as well as $\rho^{\otimes m}$ for some quantum state $\rho \in \mathbb{C}^{d\times d}$ of rank at most r, the algorithm outputs a number $\widehat{\boldsymbol{\tau}}$ and a (diagonal) estimate $\widehat{\boldsymbol{\rho}}[S] \in \mathbb{C}^{S\times S}$ such that, writing $\tau = \operatorname{tr}\rho[S]$ and recalling the notation $m_\delta = m/(c\ln(1/\delta))$ (where $c \ge 1$ is some universal constant), we have the following:

(i) if $\tau \le 1/m_\delta$ then $\widehat{\boldsymbol{\tau}} \le 1.1/m_\delta$ except with probability at most δ;

(ii) if $\widehat{\boldsymbol{\tau}} \le 1.1/m_\delta$, then $\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2 \le O(\tau \cdot f(d,r)/m)$ except with probability at most .0001;

and if $\tau \ge 1/m_\delta$ then the following hold:

(iii) the quantities τ, $\widehat{\boldsymbol{\tau}}$, and $\operatorname{tr}\widehat{\boldsymbol{\rho}}[S]$ are all within a 1.1-factor, except with probability at most δ;

(iv) $\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2 \le \tau \cdot f(d,r)/m_\delta$, except with probability at most δ;

(v) simultaneously for all $i \in S$ with $\rho_{ii} \ge \theta \coloneqq \max\{\tau/(100r), 1/m_{\delta/d}\}$, we have that $\widehat{\boldsymbol{\rho}}_{ii}$ is within a 1.1-factor of ρ_{ii}, except with probability at most δ.

(vi) simultaneously for all $i \in S$ with $\rho_{ii} \le \theta$, we have that $\widehat{\boldsymbol{\rho}}_{ii} \le 1.1\theta$, except with probability at most δ.

Proof. Since the definition of m_δ anyway contains an unspecified constant c, it is sufficient to prove the proposition with constant losses on various bounds (and then raise c's value to compensate). In particular, for notational simplicity we assume that we get $2m$ rather than m copies of ρ.

The algorithm begins by using the first m copies of ρ to obtain $(\rho_{|S})^{\otimes \boldsymbol{m}'}$ as in Proposition 3.15; this is done just to get $\boldsymbol{m}'$. The algorithm's output $\widehat{\boldsymbol{\tau}}$ is $\boldsymbol{m}'/m$, and the proposition's conclusion Item (i) follows straightforwardly from Chernoff bounds (assuming c is sufficiently large). Similarly, the τ-vs.-$\widehat{\boldsymbol{\tau}}$ part of Item (iii) follows from a Chernoff bound, and we will actually ensure 1.01-factor closeness for later convenience.

If $\widehat{\boldsymbol{\tau}} \le 1.1/m_\delta$, then the algorithm runs Proposition 3.15 on the second m copies of ρ, outputting the result. The conclusion in Item (ii) then holds except with probability at most .0001, by applying Markov's inequality to Proposition 3.15's guarantee.

We now describe how the remainder of the algorithm proceeds, when $\widehat{\boldsymbol{\tau}} \ge 1.1/m_\delta$. Note that since it only remains to prove Items (iii) to (vi), we may as well assume $\tau \ge 1/m_\delta$. The algorithm proceeds similarly to Proposition 3.15, using the second m copies of ρ to get $(\rho_{|S})^{\otimes \boldsymbol{m}'}$ for a new value of $\boldsymbol{m}'$. Since we are now assuming $\tau \ge 1/m_\delta$, a Chernoff bound implies that except with probability at most δ we'll have

$$\left|\frac{\boldsymbol{m}'}{m} - \tau\right| \le .01\sqrt{\tau/m_\delta} \le .01\tau \quad \Longrightarrow \quad \frac{\boldsymbol{m}'}{m} \text{ is within a 1.02-factor of } \tau \tag{49}$$

(as always, assuming c is large enough). The algorithm now applies Proposition 3.11 in place of Proposition 3.15, getting an estimate $\widehat{\boldsymbol{\rho}}_{|S}$ of $\rho_{|S}$ that satisfies the conclusions of Proposition 3.11. Finally, as before, the algorithm produces $\widehat{\boldsymbol{\rho}}[S] \coloneqq (\boldsymbol{m}'/m)\widehat{\boldsymbol{\rho}}_{|S}$ as its final estimate. Let us now verify Items (iii) to (vi).

First, $\operatorname{tr}\widehat{\boldsymbol{\rho}}[S] = \boldsymbol{m}'/m$, which by Equation (49) is within a 1.02-factor of τ, thereby completing the proof of Item (iii) (recall that τ and $\widehat{\boldsymbol{\tau}}$ are within a 1.01-factor).

Next, we verify Item (iv) up to a constant factor (as is sufficient). Following Equations (43) to (45) (but without expectations), we have

$$\|\rho[S] - \widehat{\boldsymbol{\rho}}[S]\|_{\mathrm{F}}^2 \le \tau^2(2\|\rho_{|S} - \widehat{\boldsymbol{\rho}}_{|S}\|_{\mathrm{F}}^2 + 2\|\boldsymbol{R}\|_{\mathrm{F}}^2) \le 2\tau^2 \cdot f(d,r)/\boldsymbol{m}'_\delta + 2(\tau - \boldsymbol{m}'/m)^2. \tag{50}$$

But Equation (49) (and using $f \ge 1$) we can bound the above by $2.03\tau \cdot f(d,r)/m_\delta$, establishing Item (iv).

To show Item (v), let B denote the set of all $i \in S$ with $\rho_{ii} \ge \theta$. Since $\theta \ge \tau/(100r) = (\operatorname{tr}\rho[S])/(100r)$, we know that $|B| \le 100r$. Moreover, for any $i \in B$ we may use

$$(\rho_{|S})_{ii} \ge \theta/\tau \ge 1/(\tau m_{\delta/d}) \ge 1/(\tau m_{\delta/r}) \ge 1/(1.02\boldsymbol{m}'_{\delta/r}), \tag{51}$$

(employing Equation (49)). (We weakened δ/d to δ/r just to illustrate this is all we need for Item (v).) So by using the second bullet point of Proposition 3.11 in a union bound over the at most $O(r)$ indices in B, we conclude that (except with probability at most $O(\delta)$) for all $i \in B$ it holds that $(\widehat{\boldsymbol{\rho}}_{|S})_{ii}$ is within a 1.01-factor of $(\rho_{|S})_{ii}$, and hence (by Equation (49)) $\widehat{\boldsymbol{\rho}}_{ii}$ is within a 1.1-factor of ρ_{ii}. This completes the verification of Item (v).

Finally, verifying Item (vi) is similar; for simplicity, we just union-bound over all $i \in S \subseteq [d]$, using the fact that $\theta \ge 1/m_{\delta/d}$. ∎

3.3 The plan for learning in χ^2: refining diagonal estimates on submatrices

Suppose we have come up with a diagonal estimate σ_1 of $\rho \in \mathbb{C}^{d\times d}$ having some Frobenius-squared distance $\eta_1 = \|\rho - \sigma_1\|_{\mathrm{F}}^2$. (Here we will have "revised" some original ρ by the unitary that makes σ_1 diagonal; this revision will be taken into account in all future uses of ρ.) Suppose we now choose some $d_2 \le d_1 \coloneqq d$, define ρ_2 to be the top-left $d_2 \times d_2$ submatrix of ρ, and apply Proposition 3.16 to it. The idea is that we hope to improve the top-left part of our estimate σ_1.

Recall that Proposition 3.16 affords us a diagonal estimate $\sigma_2 \in \mathbb{C}^{d_2\times d_2}$; let us understand a little more carefully what this means. The algorithm will give us a unitary $U_2 \in \mathbb{C}^{d_2\times d_2}$ such that $\|U_2\rho_2 U_2^\dagger - \sigma_2\|_{\mathrm{F}}^2 = \eta_2$ for some small value η_2. The idea now is to "revise" both $\rho_1 \coloneqq \rho$ and σ_1 by the unitary $U_2 \oplus \mathbb{1}$, where here $\mathbb{1}$ has dimension $d_1 - d_2$. By design, the revised version of ρ_2 will have Frobenius-squared distance η_2 from σ_2. Moreover, after revision, the fact that $\|\rho_1 - \sigma_1\|_{\mathrm{F}}^2 = \eta_1$ is unchanged (since Frobenius distance is unitarily invariant). On the other hand, although σ_1 was previously diagonal, it no longer will be after revision. But it's easy to see that it will remain diagonal *except* on its top-left $d_2 \times d_2$ block, which we are intending to replace by σ_2 anyway. In particular, the *off-diagonal* $d_2 \times (d_1 - d_2)$ and $(d_1 - d_2) \times d_2$ blocks of σ_1 remain zero.

Let us summarize. We will first obtain a diagonal estimate σ_1 of ρ_1 with some error η_1. Then after choosing some $d_2 \in [d_1]$, we will obtain a further diagonal estimate σ_2 of the top-left $d_2 \times d_2$ block of ρ, with some error η_2. We might then take as final estimate $\widehat{\rho}$ the diagonal matrix formed by replacing the top-left $d_2 \times d_2$ block of σ_1 by σ_2.

Naturally, this plan can be iterated (meaning we can try to improve the estimate's top-left $d_3 \times d_3$ block for some $d_3 \in [d_2]$) but let us pause here to discuss error. If we're interested in the Frobenius-squared error of our current estimate $\widehat{\rho}$, we can't say more than that it is bounded by $\eta_1 + \eta_2$. Here we're decomposing the error into the contribution from the top-left $d_2 \times d_2$ block (which is η_2) plus the contribution from the remaining ⌟-shaped region (consisting of the bottom-right $(d_1 - d_2) \times (d_1 - d_2)$ block plus the two off-diagonal blocks). We will just bound this second error contribution by the whole Frobenius-squared distance of σ_1 from ρ_1, which is η_1.

It would seem that this scheme of refining our estimate for the top-left block hasn't helped, since it took us from Frobenius-squared error η_1 to Frobenius-squared error (at most) $\eta_1 + \eta_2$. But the idea is that our new estimate $\widehat{\rho}$ may have improved *Bures χ^2-divergence*. Recall the formula for χ^2-divergence, Equation (33) (which we will apply even though $\widehat{\rho}$ might not precisely be a state, meaning of trace 1). Recall also that our diagonal estimates $\sigma_1 = \mathrm{diag}(q^{(1)})$ and $\sigma_2 = \mathrm{diag}(q^{(2)})$ are chosen to have nondecreasing entries along the diagonal. (We moreover expect that $\widehat{\rho}$ will also have nondecreasing entries, meaning $q^{(2)}_{d_2} \le q^{(1)}_{d_2+1}$, but we won't rely on this.) Now we can use the bound

$$\mathrm{D}_{\chi^2}(\rho \,\|\, \widehat{\rho}) \le \sum_{i,j=1}^{d} \frac{2}{q_{\max(i,j)}} |\rho_{ij} - \widehat{\rho}_{ij}|^2 \le \frac{2}{q^{(2)}_1} \sum_{i,j=1}^{d_2} |\rho_{ij} - \widehat{\rho}_{ij}|^2 + \frac{2}{q^{(1)}_{d_2+1}} \sum_{\substack{i,j:\\ \max(i,j)>d_2}} |\rho_{ij} - \widehat{\rho}_{ij}|^2$$
$$\le \frac{2}{q^{(2)}_1}\eta_2 + \frac{2}{q^{(1)}_{d_2+1}}\eta_1. \quad (52)$$

The idea here is that if, perhaps

$$q^{(2)}_1 \approx \cdots q^{(2)}_{d_2} \approx (\mathrm{tr}\,\sigma_2)/r; \quad \text{and} \quad q^{(1)}_{d_2+1} \approx \cdots \approx q^{(1)}_d \approx (\mathrm{tr}\,\sigma_1)/r, \quad (53)$$

then hopefully from Proposition 3.15 with m copies we will have $\eta_1 \approx O(\tau_1 \cdot f(d,r))/m$ and $\eta_2 \approx O(\tau_2 \cdot f(d,r))/m$, where $\tau_i = \mathrm{tr}\,\rho_i \approx \mathrm{tr}\,\sigma_i$. Then the total χ^2-error would be approximately

$$\frac{2r}{\tau_2} \cdot O(\tau_2 \cdot f(d,r))/m + \frac{2r}{\tau_1} \cdot O(\tau_1 \cdot f(d,r))/m = O(r \cdot f(d,r))/m. \quad (54)$$

This would mean we have converted Frobenius-squared rate $O(f(d,r))$ to χ^2-divergence rate $O(r \cdot f(d,r))$. Now Equation (53) might seem a little optimistic, but our idea will be that no matter what ρ's eigenvalues are, we can break them up into logarithmically many groups where they only differ by a constant factor, and thereby achieve the desired χ^2-divergence rate of $O(r \cdot f(d,r))$ up to logarithmic losses. Unfortunately, we will have to deal separately with any extremely small eigenvalues of ρ, which causes some additional losses.

Remark 3.17. Ideally this plan suggests we might be able to achieve sample complexity $n = \widetilde{O}(rd/\epsilon)$ for tomography with respect to Bures χ^2-divergence (for collective measurements). But the "small eigenvalue" issue causes problems for this. Without explicitly claiming a lower bound, let us sketch why it seems difficult to significantly beat the $n = \widetilde{O}(r^{.5}d^{1.5}/\epsilon)$ complexity from Corollary 1.7, even in the case $r = 1$.

So suppose $\rho = |v\rangle\langle v|$ for an unknown unit vector $|v\rangle \in \mathbb{C}^d$. The best known tomography algorithm for a pure state is extremely natural and simple [24]; it outputs a pure state $|\boldsymbol{u}\rangle\langle\boldsymbol{u}|$ and achieves $|\langle \boldsymbol{u}|\boldsymbol{v}\rangle| \geq 1 - \eta$, i.e. infidelity η, with high probability using $n = O(d/\eta)$ copies. Moreover, $n = \Omega(d/\eta)$ is a known lower bound [21]. However, with certainty we will have $|\langle \boldsymbol{u}|\boldsymbol{v}\rangle| \neq 1$ and hence $\mathrm{D}_{\chi^2}(\rho \parallel |\boldsymbol{u}\rangle\langle\boldsymbol{u}|) = \infty$. To achieve $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) < \infty$ we will have to output a full-rank hypothesis $\widehat{\boldsymbol{\rho}}$, and to achieve $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \leq O(\epsilon)$ it's hard to imagine what to try besides something like $\widehat{\boldsymbol{\rho}} = \Delta_\epsilon(|\boldsymbol{u}\rangle\langle\boldsymbol{u}|)$. But with this choice it's not hard to compute that $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \geq (d/\epsilon) \cdot \Omega(\eta^2)$, seemingly forcing us to choose $\eta = \Theta(\epsilon/\sqrt{d})$ and thereby use $n = \Omega(d^{1.5}/\epsilon)$ copies.

3.4 The central estimation algorithm

Theorem 3.18. *A state estimation algorithm $\mathcal{A}$ having Frobenius-squared rate $f = f(d, r)$ satisfying*[4] *$f \gg \log d$ may be transformed (preserving the single-copy measurement property) into a state estimation algorithm $\mathcal{A}'$ returning diagonal estimates with the following properties:*

Given r and $m \geq r$, the algorithm sets the following parameters:

$$\delta = \frac{.0001}{\log_2(m/rf)}, \qquad \tilde{\epsilon} = Crf/m_\delta, \qquad \ell_{\max} = \lceil \log_2(1/\tilde{\epsilon}) \rceil, \qquad \epsilon = \tilde{\epsilon}\ell_{\max} \qquad M = 2m\ell_{\max}.$$

(Here C is a large universal constant, and it is assumed that ϵ is at most some small universal constant.)

Then, given $\rho^{\otimes M}$, where $\rho \in \mathbb{C}^{d\times d}$ is a quantum state of rank at most r, the algorithm $\mathcal{A}'$ outputs (with probability at least .99) a partition $[d] = \boldsymbol{L} \sqcup \boldsymbol{R}$ (with $\boldsymbol{L} = [\boldsymbol{d}']$ for some $\boldsymbol{d}' \leq d$), together with a quantum state $\widetilde{\boldsymbol{\rho}} = \mathrm{diag}(\boldsymbol{q}) \in \mathbb{C}^{d\times d}$ satisfying:

(a) $|\boldsymbol{R}| \leq O(r\ell_{\max})$;

(b) $\boldsymbol{\tau}, \boldsymbol{\epsilon}' \leq O(\tilde{\epsilon})$, *where* $\boldsymbol{\tau} \coloneqq \mathrm{tr}\,\rho[\boldsymbol{L}]$ *and* $\boldsymbol{\epsilon}' \coloneqq \mathrm{tr}\,\widetilde{\boldsymbol{\rho}}[\boldsymbol{L}]$;

(c) $\|\rho[\boldsymbol{L}] - \widetilde{\boldsymbol{\rho}}[\boldsymbol{L}]\|_{\mathrm{F}}^2 \leq O(\frac{\tilde{\epsilon}^2}{r\ln(1/\delta)}) \leq O(\tilde{\epsilon}^2/r)$;

(d) $\widehat{\mathrm{D}}_{\chi^2}^{-\boldsymbol{L}}(\rho \parallel \widetilde{\boldsymbol{\rho}}) \leq O(\epsilon)$.

Proof. Fix a Frobenius-squared estimation algorithm $\mathcal{A}$ with rate $f = f(d, r)$, and assume we have passed it through Proposition 3.16 so that we may use it to make diagonal estimates of subnormalized states.

The algorithm $\mathcal{A}'$ will run in some $\boldsymbol{\ell}$ stages, where we guarantee $\boldsymbol{\ell} \leq \ell_{\max}$. Each stage will consume m copies of ρ. After the $\boldsymbol{\ell}$th stage, there will be some final processing that uses the remaining $M/2$ (at least) copies of ρ.

As the algorithm progresses, it will define a sequence of numbers $d = \boldsymbol{d}_1 \geq \boldsymbol{d}_2 \geq \cdots \geq \boldsymbol{d}_{\boldsymbol{\ell}}$, with the value $\boldsymbol{d}_{t+1}$ being selected at the end of the tth stage. We introduce the notation $\boldsymbol{R}_t = \{\boldsymbol{d}_{t+1} + 1, \ldots, \boldsymbol{d}_t\}$; each of these sets will have cardinality at most r.

At the beginning of the tth stage, $\mathcal{A}'$ will run the algorithm from Proposition 3.16 on $\rho[\boldsymbol{d}_t]$, with confidence parameter δ, resulting in some $\widehat{\boldsymbol{\tau}}_t$ and a diagonal estimate that we will call $\boldsymbol{\sigma}_t$. We will use the fact that δ always satisfies all of the following (provided C is large enough and using $f \geq \log d$):

$$1/m_\delta \leq \tilde{\epsilon}, \quad 1/m_{\delta/d} \ll \tilde{\epsilon}/r, \quad \delta \leq .0001/\ell_{\max}. \tag{55}$$

By losing probability at most 5δ in each stage, we may assume that except with probability at most .0006, all of the desired outcomes from Proposition 3.16 do occur over the course of the algorithm.

If $\widehat{\boldsymbol{\tau}}_t \leq 1.1\tilde{\epsilon}$ or $t > d$, then this is declared the final stage; i.e., the algorithm will define $\boldsymbol{\ell} = t$ and move to its "final processing". Otherwise, in a non-final stage we have $\widehat{\boldsymbol{\tau}} > 1.1\tilde{\epsilon} \geq 1.1/m_\delta$ (using Inequality (55)), so by Item (i) of Proposition 3.16 we may assume that $\mathrm{tr}\,\rho[\boldsymbol{d}_t] > 1/m_\delta$, and hence (using Item (iii)),

$$\text{for } t < \boldsymbol{\ell}, \qquad \mathrm{tr}\,\boldsymbol{\sigma}_t[S] \text{ is within a 1.1-factor of } \tau_t \coloneqq \mathrm{tr}\,\rho[\boldsymbol{d}_t]; \qquad \text{moreover, } \tau_t \geq \tilde{\epsilon}. \tag{56}$$

[4] This mild assumption is made to keep parameter-setting simpler.

Next, using Item (v) of Proposition 3.16 and $1/m_{\delta/d} \ll \tilde{\epsilon}/r \le \tau_t/r$ (which implies that the Proposition's "θ" is $\tau/(100r)$), we get that

$$\text{for } t < \boldsymbol{\ell}, \quad \text{for all } i \le \boldsymbol{d}_t \text{ with } \rho_{ii} \ge \tau_t/(100r), \quad (\boldsymbol{\sigma}_t)_{ii} \text{ is within a 1.1-factor of } \rho_{ii}. \tag{57}$$

Moreover, from Item (vi) we get

$$\text{for } t < \boldsymbol{\ell}, \quad \text{for all } i \le \boldsymbol{d}_t \text{ with } \rho_{ii} \le \tau_t/(100r), \quad (\boldsymbol{\sigma}_t)_{ii} \le 1.1\tau_t/(100r). \tag{58}$$

Finally, we record the main conclusions Items (ii) and (iv) of Proposition 3.16, taking care to distinguish the final stage:

$$\text{for } t < \boldsymbol{\ell}, \quad \|\rho[\boldsymbol{d}_t] - \boldsymbol{\sigma}_t\|_{\mathrm{F}}^2 \le \tau_t f/m_\delta; \qquad \text{and} \qquad \text{for } t = \boldsymbol{\ell}, \quad \|\rho[\boldsymbol{d}_t] - \boldsymbol{\sigma}_t\|_{\mathrm{F}}^2 \le O(\tau_{\boldsymbol{\ell}} f/m). \tag{59}$$

Now we explain how algorithm $\mathcal{A}'$ defines $\boldsymbol{d}_{t+1}$ at the end of non-final stage t, where recall non-finality implies from Inequality (56)

$$\tau_t \ge \tilde{\epsilon} = Crf/m_\delta. \tag{60}$$

Considering the first bound in Inequality (59), note that $\operatorname{rank} \rho[\boldsymbol{d}_t] \le r$, so we have that the diagonal matrix $\boldsymbol{\sigma}_t$ has Frobenius-squared distance at most $\tau_t f/m_\delta$ from a matrix of rank at most r. But the rank-at-most-r matrix that is Frobenius-squared-closest to $\boldsymbol{\sigma}_t$ is simply $\boldsymbol{\sigma}'_t$, the matrix formed by zeroing out all but the r largest entries of $\boldsymbol{\sigma}_t$. Recalling that $\boldsymbol{\sigma}'_t$ has nondecreasing diagonal entries, this means $\boldsymbol{\sigma}'_t$ is formed by zeroing out all diagonal entries of index at most $\boldsymbol{d}'_{t+1} \coloneqq \max\{\boldsymbol{d}_t - r, 0\}$. Thus we have

$$\|\rho[\boldsymbol{d}_t] - \boldsymbol{\sigma}'_t\|_{\mathrm{F}}^2 \le 4\tau_t f/m_\delta \quad \Longrightarrow \quad \|\rho[\boldsymbol{d}'_{t+1}]\|_{\mathrm{F}}^2 \le 4\tau_t f/m_\delta \quad \Longrightarrow \quad (\operatorname{tr} \rho[\boldsymbol{d}'_{t+1}])^2 \le r \cdot 4\tau_t f/m_\delta, \tag{61}$$

where the last deduction used $\operatorname{rank} \rho[\boldsymbol{d}'_{t+1}] \le r$. But assuming $C \ge 64$, Inequality (60) implies

$$r \cdot 4\tau_t f/m_\delta = \tau_t \cdot (4rf/m_\delta) \le \tau_t \cdot \tfrac{1}{16}\tau_t = (\tfrac{1}{4}\tau_t)^2, \tag{62}$$

Thus from Inequality (61) we conclude

$$\operatorname{tr} \rho[\boldsymbol{d}'_{t+1}] \le \tfrac{1}{4}\tau_t \quad \Longrightarrow \quad \operatorname{tr} \rho[\boldsymbol{R}'_t] \ge \tfrac{3}{4}\tau_t \text{ for } \boldsymbol{R}'_t \coloneqq \{\boldsymbol{d}'_{t+1} + 1, \dots, \boldsymbol{d}_t\}. \tag{63}$$

Let $\boldsymbol{R}''_t = \{i \in \boldsymbol{R}'_t : \rho_{ii} > (1.1)^4\tau_t/(100r) = .014641\tau_t/r\}$. Since $\left|\boldsymbol{R}'_t\right| \le r$, the sum of ρ_{ii} over all $\boldsymbol{R}'_t \setminus \boldsymbol{R}''_t$ is at most $.014641\tau_t \le .02\tau_t$; hence

$$\sum_{i \in \boldsymbol{R}''_{t+1}} \rho_{ii} \ge .73\tau_t, \quad \text{and each summand exceeds } (1.1)^4 \frac{\tau_t}{100r}. \tag{64}$$

Applying Inequalities (56) and (57), we conclude that

$$\sum_{i \in \boldsymbol{R}''_t} (\boldsymbol{\sigma}_t)_{ii} \ge \tfrac{.73}{1.1}\tau_t > .66\tau_t, \quad \text{and each summand exceeds } (1.1)^3 \frac{\tau_t}{100r} \ge (1.1)^2 \frac{\operatorname{tr} \boldsymbol{\sigma}_t}{100r}. \tag{65}$$

We now stipulate that algorithm $\mathcal{A}'$ chooses $\boldsymbol{d}_{t+1} \ge \boldsymbol{d}'_{t+1}$ to be minimal so that

$$(\boldsymbol{\sigma}_t)_{ii} > (1.1)^2 \frac{\operatorname{tr} \boldsymbol{\sigma}_t}{100r} \quad \text{for all } i \in \boldsymbol{R}_t = \{\boldsymbol{d}_{t+1} + 1, \dots, \boldsymbol{d}_t\}. \tag{66}$$

(In other words, $\{\boldsymbol{d}_{t+1} + 1, \dots, \boldsymbol{d}_t\}$ is the maximum-cardinality suffix of $\boldsymbol{R}'_t$ where the above holds.) Then

$$\sum_{i \in \boldsymbol{R}_t} (\boldsymbol{\sigma}_t)_{ii} = \operatorname{tr} \boldsymbol{\sigma}_t[\boldsymbol{R}_t] > .66\tau_t. \tag{67}$$

Moreover, from Inequalities (56) and (58) we know that $\rho_{ii} > \tau_t/(100r)$ for all $i \in \boldsymbol{R}_t$; hence

$$(\sigma_t)_{ii} \text{ is within a 1.1-factor of } \rho_{ii} \text{ for all } i \in \boldsymbol{R}_t, \tag{68}$$

and therefore Inequality (67) implies $\operatorname{tr}\rho[\boldsymbol{R}_t] > \frac{.66}{1.1}\tau_t \geq \frac{1}{2}\tau_t$, whence

$$\tau_{t+1} = \operatorname{tr}\rho[\boldsymbol{d}_{t+1}] = \operatorname{tr}\rho[\boldsymbol{d}_t] - \operatorname{tr}\rho[\boldsymbol{R}_t] \leq \tau_t - \tfrac{1}{2}\tau_t < \tfrac{1}{2}\tau_t. \tag{69}$$

This is the key deduction that lets us make progress, in particular confirming that $\boldsymbol{\ell} \leq \lceil \log_2(1/\tilde{\epsilon}) \rceil$ (because of Inequality (56)).

Now we discuss the "final processing". The final partition output by $\mathcal{A}'$ will be $\boldsymbol{L} \sqcup \boldsymbol{R}$, where $\boldsymbol{L} = [\boldsymbol{d}_{\boldsymbol{\ell}}]$ and $\boldsymbol{R} = \boldsymbol{R}_1 \sqcup \cdots \sqcup \boldsymbol{R}_{\boldsymbol{\ell}-1}$; since $|\boldsymbol{R}_t| \leq r$ for all t we satisfy the theorem's Item (a). We can verify the bound on $\boldsymbol{\tau} = \operatorname{tr}\rho[\boldsymbol{d}_{\boldsymbol{\ell}}]$ in Item (b) by recalling that when the final stage is reached we have $\widehat{\boldsymbol{\tau}} \leq 1.1\tilde{\epsilon}$, and hence $\boldsymbol{\tau} \leq (1.1)^2\tilde{\epsilon}$ by Item (iii) of Proposition 3.16 (recall $\tilde{\epsilon} \geq 1/m_\delta$). We can also partly verify the conclusion Item (c) by observing that the *off-diagonal* Frobenius-squared of $\rho[\boldsymbol{L}]$ is upper-bounded by $\|\rho[\boldsymbol{L}] - \boldsymbol{\sigma}_{\boldsymbol{d}_{\boldsymbol{\ell}}}\|_{\mathrm{F}}^2$, and by Inequality (59) this is at most $O(\tau_{\boldsymbol{\ell}} f/m) \leq O(\tilde{\epsilon} f/m) = O(\tilde{\epsilon}^2)/(r\ln(1/\delta))$. Thus:

$$\text{Item (c) holds provided } \|\operatorname{diag}(\rho)[\boldsymbol{L}] - \widetilde{\boldsymbol{\rho}}[\boldsymbol{L}]\|_2^2 \leq O(\tilde{\epsilon}^2)/(r\ln(1/\delta)). \tag{70}$$

Aside from establishing the above, it remains to describe how algorithm $\mathcal{A}'$ forms $\widetilde{\boldsymbol{\rho}}$ satisfying the theorem's conclusion Item (d). We first describe a candidate output we'll call $\boldsymbol{\sigma}'$ that *almost* works: namely, $\boldsymbol{\sigma}'$ is formed by setting its diagonal elements from $\boldsymbol{R}_t$ to be those from $\boldsymbol{\sigma}_t$, for $t < \boldsymbol{\ell}$. (The remaining diagonal entries may be set to 0.) The difficulty with this is that it's not easy to control $\operatorname{tr}\boldsymbol{\sigma}'$, but let us ignore this issue and calculate χ^2-divergence.[5] Ignoring the fact that we are not working with normalized states, we may bound

$$\widehat{\mathrm{D}}_{\chi^2}^{-\boldsymbol{L}}(\rho \parallel \boldsymbol{\sigma}') \leq \sum_{\substack{i,j\in[d] \\ k:=\max(i,j)\in\boldsymbol{R}}} \frac{2}{\boldsymbol{\sigma}'_{kk}}\left|\rho_{ij} - \boldsymbol{\sigma}'_{ij}\right|^2 \leq \sum_{t<\boldsymbol{\ell}} \frac{2}{\min\{\boldsymbol{\sigma}'_{hh} : h \in \boldsymbol{R}_t\}} \cdot \|\rho[\boldsymbol{d}_t] - \boldsymbol{\sigma}_t\|_{\mathrm{F}}^2. \tag{71}$$

Each summand above can be upper-bounded using Inequalities (59) and (66), yielding

$$\widehat{\mathrm{D}}_{\chi^2}^{-\boldsymbol{L}}(\rho \parallel \boldsymbol{\sigma}') \leq \sum_{t<\boldsymbol{\ell}} \frac{2}{1.1\tau_t/(100r)} \cdot \tau_t f/m_\delta \leq O(\ell_{\max} r f/m_\delta) = O(\tilde{\epsilon}\ell_{\max}) = O(\epsilon). \tag{72}$$

We now work to control the trace of our estimate. Our strategy is to have $\mathcal{A}'$ perform diagonal measurements on the remaining $M/2$ copies of ρ to classically relearn its diagonal via Proposition 2.16, with its "S" set to $\boldsymbol{R}$. Calling the resulting probability distribution $\boldsymbol{q}$, the algorithm will finally take $\widetilde{\boldsymbol{\rho}} = \operatorname{diag}(\boldsymbol{q})$.

First we complete the verification by Item (c) by establishing the condition in Inequality (70): since $\boldsymbol{q}[\boldsymbol{L}]$ is formed by the empirical estimator, Markov's inequality and Proposition 2.14 imply that except with probability at most .0001 we have $\|\operatorname{diag}(\rho)[\boldsymbol{L}] - \widetilde{\boldsymbol{\rho}}[\boldsymbol{L}]\|_2^2 \leq O(\operatorname{tr}\rho[\boldsymbol{L}])/(M/2) \leq O(\tilde{\epsilon}/(m\ell_{\max})) = O(\tilde{\epsilon}^2/(r\ln(1/\delta) f\ell_{\max})$, and we have a factor of $f\ell_{\max}$ to spare.

Next, using Markov again with Proposition 2.16 we get that except with probability at most .0001,

$$\mathrm{d}_{\chi^2}(\operatorname{diag}(\rho[\boldsymbol{R}]) \parallel \boldsymbol{q}[\boldsymbol{R}]) \leq O(|\boldsymbol{R}|/M) \leq O(r/m) \ll \tilde{\epsilon}. \tag{73}$$

Also, using $f \geq \log d$ (and C large enough), we indeed have $1/(M/2)_{\delta'} \leq \tilde{\epsilon}/(100r)$ for $\delta' = .0001/|\boldsymbol{R}| \geq .0001/(r\ell_{\max})$; since also $\rho_{ii} \geq \tilde{\epsilon}/(100r)$ for all $i \in \boldsymbol{R}$ (recall Inequality (64)), we conclude

$$\boldsymbol{q}_i \text{ is within a 4-factor of } \rho_{ii} \text{ for all } i \in \boldsymbol{R}, \tag{74}$$

except with probability at most .0001. Finally, it is easy to see that in Proposition 2.16 we have $\mathbf{E}[\|\boldsymbol{q}[\boldsymbol{L}]\|_1] \leq \operatorname{tr}\rho[\boldsymbol{L}] = \boldsymbol{\tau} \leq O(\tilde{\epsilon})$, and hence Markov implies that except with probability at most .0001 we have $\boldsymbol{\epsilon}' = \|\boldsymbol{q}[\boldsymbol{L}]\|_1 \leq O(\tilde{\epsilon})$, completing the verification of Item (b).

Finally we finish the analysis of $\widehat{\mathrm{D}}_{\chi^2}^{-\boldsymbol{L}}(\rho \parallel \widetilde{\boldsymbol{\rho}})$. The contribution to this quantity from the diagonal entries is precisely Inequality (73). On the other hand, since $\boldsymbol{\sigma}'$ and $\widetilde{\boldsymbol{\rho}}$ are both diagonal, the off-diagonal contribution to $\widehat{\mathrm{D}}_{\chi^2}^{-\boldsymbol{L}}(\rho \parallel \widetilde{\boldsymbol{\rho}})$ can be bounded by a constant times Inequality (72), using the fact that the diagonal entries (from $\boldsymbol{R}$) of $\boldsymbol{\sigma}'$ and $\widetilde{\boldsymbol{\rho}}$ are all within a constant factor by virtue of Inequalities (68) and (74). This completes the verification of Item (d). □

[5] Strictly speaking, $\boldsymbol{\sigma}'$ need not have nondecreasing diagonal entries as promised, but we can finally "revise" by a permutation matrix to fix this.

We also show the following, to improve some $\log(1/\epsilon)$ factors in the case that ϵ is extremely small and $r = \Theta(d)$. (The reader might like to think of the case when $d = O(1)$.)

Theorem 3.19. *There is a variant version of $\mathcal{A}'$ from Theorem 3.18 with the following alternative parameter settings:*

$$\delta = \frac{.0001}{d+1}, \qquad \ell_{\max} = d+1, \qquad \epsilon = \frac{d \ln r}{r}\tilde{\epsilon}. \tag{75}$$

Proof. Besides verifying that Inequality (55) still holds with our changed δ and $\ell_{\max}$, there is one alternative idea to be explained. In the preceding proof, the driver of progress was Inequality (69) showing $\tau_{t+1} < \frac{1}{2}\tau_t$; this enabled us to take $\ell_{\max}$ logarithmic in $1/\tilde{\epsilon}$. In this variant, we will only use this inequality weakly, to show that $|\boldsymbol{R}_t| \geq 1$ so that $d_{t+1} < d_t$ strictly; this is already enough to ensure that taking $\ell_{\max} = d+1$ is acceptable. On the other hand, if we only implement this change then ϵ would become unnecessarily large (namely, $\tilde{\epsilon}(d+1)$).

To get the improved value of ϵ, we change how $\mathcal{A}'$ chooses the $\boldsymbol{d}_t$ values. Returning to Inequality (65), in the tth stage there is a set $\boldsymbol{R}''_t$ of at most r indices i on which each $(\boldsymbol{\sigma}_t)_{ii}$ exceeds $\boldsymbol{\beta} \coloneqq (1.1)^2 \cdot \frac{\operatorname{tr}\boldsymbol{\sigma}_t}{100r}$, and their sum $\boldsymbol{s}$ exceeds $.66\tau_t \geq .6(\operatorname{tr}\boldsymbol{\sigma}_t)$. Then $\mathcal{A}'$ chooses $\boldsymbol{R}_t$ to consist of all indices $i \in \boldsymbol{R}'_t$ with $(\boldsymbol{\sigma}_t)_{ii} \geq \boldsymbol{\beta}$, of which there are at most $O(r)$. Note that if we conversely had $|\boldsymbol{R}_t|$ at *least* $\Omega(r)$ for every t, then the algorithm would halt in at most $O(d/r)$ stages, allowing us to take $\epsilon = (d/r)\tilde{\epsilon}$ rather than $\tilde{\epsilon}\ell_{\max}$ (a significant improvement when $r = \Theta(d)$).

The idea is now for $\mathcal{A}'$ to choose a slightly different $\boldsymbol{R}_t$ in each round, of cardinality $\boldsymbol{r}_t \geq 1$, so that $(\boldsymbol{\sigma}_t)_{ii} \geq \Omega(\frac{\operatorname{tr}\boldsymbol{\sigma}_t}{\boldsymbol{r}_t \ln r})$. (Note that we need not be concerned with the sum of $(\boldsymbol{\sigma}_t)_{ii}$ on the new $\boldsymbol{R}_t$, since we're now only using that $|\boldsymbol{R}_t| \geq 1$ always.) If we can show this is possible, then we can use it as a replacement for Inequality (66) when deriving Inequality (72); we'll then get

$$\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \boldsymbol{\sigma}') \leq \sum_{t<\boldsymbol{\ell}} \frac{O(\boldsymbol{r}_t \ln r)}{\tau_t} \cdot \tau_t f/m_\delta \leq O(d(\ln r)f/m_\delta) = O(\epsilon), \tag{76}$$

as claimed.

But the proof that we can choose $\boldsymbol{R}_t$ as described is elementary. Essentially, the algorithm has a nonincreasing sequence of (at most) r numbers $x_1, \dots, x_r$ (where $x_i = (\boldsymbol{\sigma}_t)_{\boldsymbol{d}_t+1-i,\boldsymbol{d}_t+1-i}$) whose sum is (at least) $\boldsymbol{s}$. We need to show that for some $\boldsymbol{r}_t$ it holds that $x_{\boldsymbol{r}_t} \geq \Omega(\frac{\boldsymbol{s}}{\boldsymbol{r}_t \ln r})$. But if $x_k \ll \frac{\boldsymbol{s}}{k \ln r}$ for all $k \in [r]$, then $\sum_{k=1}^{r} x_k \ll \boldsymbol{s}$, a contradiction. □

3.5 Conclusions from the central estimation algorithm

Corollary 3.20. *After applying Theorem 3.18 and introducing the quantum state $\widehat{\boldsymbol{\rho}} = \widetilde{\boldsymbol{\rho}}_{|\boldsymbol{R}} = \frac{1}{1-\boldsymbol{\epsilon}'}\widetilde{\boldsymbol{\rho}}[\boldsymbol{R}]$ (extended with 0's so it is in $\mathbb{C}^{d\times d}$), we have*

$$\mathrm{D}^2_{\mathrm{B}}(\rho, \widehat{\boldsymbol{\rho}}) \leq O(\epsilon). \tag{77}$$

Proof. Let us write $\rho_{|\boldsymbol{R}} = \frac{1}{1-\boldsymbol{\tau}}\boldsymbol{\rho}[\boldsymbol{R}]$ (which again we'll extend to $\mathbb{C}^{d\times d}$ when necessary). Let us first show

$$\mathrm{D}_{\chi^2}(\rho_{|\boldsymbol{R}} \parallel \widetilde{\boldsymbol{\rho}}_{|\boldsymbol{R}}) \leq O(\epsilon). \tag{78}$$

Up to a slight "rescaling" by the factors $1-\boldsymbol{\tau}$ and $1-\boldsymbol{\epsilon}'$, this is nearly the same as the Item (d) conclusion,

$$\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widetilde{\boldsymbol{\rho}}) \leq O(\epsilon). \tag{79}$$

Item (b) tells us that $\boldsymbol{\tau}, \boldsymbol{\epsilon}' \leq O(\tilde{\epsilon})$ (which may be assumed at most, say, $\frac{1}{2}$); from this, it is not hard to show that the "rescaling" only makes a constant-factor difference to the off-diagonal χ^2-divergence contributions. So to establish Inequality (78), it remains to analyze the effect of rescaling on the on-diagonal χ^2-divergence contributions. Writing $\rho_{ii} = (1+\boldsymbol{\zeta}_i)\boldsymbol{q}_i$ for some numbers $\boldsymbol{\zeta}_i > 0$, the bound on just the diagonal contribution in Inequality (79) is equivalent to

$$\sum_{i\in\boldsymbol{R}} \boldsymbol{\zeta}_i^2 \boldsymbol{q}_i \leq O(\epsilon). \tag{80}$$

Then in the rescaling, when ρ_{ii} is replaced by $\frac{1}{1-\boldsymbol{\tau}}\rho_{ii}$ in $\rho_{|\boldsymbol{R}}$ and $\boldsymbol{q}_i$ is replaced by $\frac{1}{1-\boldsymbol{\epsilon}'}\boldsymbol{q}_i$ in $\widehat{\rho}$, it is as though $\boldsymbol{\zeta}_i$ is replaced by $\frac{1-\boldsymbol{\epsilon}'}{1-\boldsymbol{\tau}}\boldsymbol{\zeta}_i = (1 \pm O(\tilde{\boldsymbol{\epsilon}}))\boldsymbol{\zeta}_i$. Putting that into Inequality (80) shows that the rescaling only changes the on-diagonal χ^2-divergence contribution by an additive $O(\tilde{\epsilon})\sum_{i\in\boldsymbol{R}}\boldsymbol{q}_i = O(\tilde{\epsilon})$, sufficient to complete the proof of Inequality (78).

Having established Inequality (78) (and recalling $\widehat{\boldsymbol{\rho}} = \widetilde{\boldsymbol{\rho}}_{|\boldsymbol{R}}$), Theorem 2.32 immediately implies

$$\mathrm{D}_{\mathrm{B}}^2(\rho_{|\boldsymbol{R}}, \widehat{\boldsymbol{\rho}}) \le O(\epsilon). \tag{81}$$

On the other hand, the Gentle Measurement Lemma [49]

$$\operatorname{tr}\rho[\boldsymbol{R}] = \mathrm{F}(\rho, \rho_{|\boldsymbol{R}})^2 = (1 - \tfrac{1}{2}\mathrm{D}_{\mathrm{B}}^2(\rho, \rho_{|\boldsymbol{R}}))^2. \tag{82}$$

From $\operatorname{tr}\rho[\boldsymbol{R}] = 1 - \boldsymbol{\tau} \ge 1 - O(\tilde{\epsilon})$, the above directly yields $\mathrm{D}_{\mathrm{B}}^2(\rho, \rho_{|\boldsymbol{R}}) \le O(\tilde{\epsilon})$, and thus Inequality (77) follows from Inequality (81) and $\mathrm{D}_{\mathrm{B}}(\cdot,\cdot)$ being a metric. □

Now by working out the parameters (using both Theorems 3.18 and 3.19), we get the below Frobenius-to-infidelity transformation. The further transformation to relative entropy accuracy promised in Theorem 1.6 follows by applying Theorem 2.34.

Corollary 3.21. *A state estimation algorithm with Frobenius-squared rate $f = f(d,r) \gg \log d$ may be transformed (preserving the single-copy measurement property) into a state estimation algorithm with the following property:*

Given parameters ϵ, r, and M copies of a quantum state $\rho \in \mathbb{C}^{d\times d}$ of rank at most r, either

$$M = O\left(\frac{rf(d,r)}{\epsilon}\cdot\log^2(1/\epsilon)\log\log(1/\epsilon)\right) \quad \textit{or alternatively,} \quad M = O\left(\frac{1}{\epsilon}\cdot d^2 f(d,r)(\log d)(\log r)\right), \tag{83}$$

suffices for the algorithm to output (with probability at least .99) the classical description of a quantum state $\widehat{\boldsymbol{\rho}}$ with infidelity $1 - \mathrm{F}(\rho, \widehat{\boldsymbol{\rho}}) = \frac{1}{2}\mathrm{D}_{\mathrm{B}}^2(\rho, \widehat{\boldsymbol{\rho}}) \le \epsilon$.

In particular, by Theorem 1.3

$$M = O\left(\frac{rd}{\epsilon}\cdot\log^2(1/\epsilon)\log\log(1/\epsilon)\right) \textit{ suffices using collective measurements} \tag{84}$$

(or for very small d, alternatively $M = O\big(\frac{1}{\epsilon}\cdot d^3(\log d)(\log r)\big)$ suffices). And, by Theorem 1.1

$$M = O\left(\frac{r^2d}{\epsilon}\cdot\log^2(1/\epsilon)\log\log(1/\epsilon)\right) \textit{ suffices using single-copy measurements} \tag{85}$$

(or for very small d, alternatively $M = O\big(\frac{1}{\epsilon}\cdot rd^3(\log d)(\log r)\big)$).

It remains to obtain the Frobenius-to-χ^2 transformation promised in Theorem 1.6. This is Corollary 3.24 below, which we achieve in two steps.

Corollary 3.22. *After applying Theorem 3.18 and introducing the quantum state $\widehat{\boldsymbol{\rho}} = \eta\cdot\frac{1}{\boldsymbol{L}}\mathbb{1}_{\boldsymbol{L}\times\boldsymbol{L}} + (1-\eta)\widetilde{\boldsymbol{\rho}}$ (where we assume $\eta < 1/2$, say) we have*

$$\mathrm{D}_{\chi^2}^{\mathrm{off}}(\rho \parallel \widehat{\boldsymbol{\rho}}) \le O(\tilde{\epsilon}\log(1/\tilde{\epsilon}) + (d/r)(\tilde{\epsilon}^2/\eta)) \tag{86}$$

$$\mathrm{D}_{\chi^2}^{\mathrm{on}}(\rho \parallel \widehat{\boldsymbol{\rho}}) \le O(\eta + \tilde{\epsilon}\log(1/\tilde{\epsilon}) + (d/r)(\tilde{\epsilon}^2/\eta)) \tag{87}$$

$$\implies \quad \mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \le O(\eta + \tilde{\epsilon}\log(1/\tilde{\epsilon}) + (d/r)(\tilde{\epsilon}^2/\eta)), \tag{88}$$

where we have split out the "off-diagonal" and "on-diagonal" contributions to $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}})$. (Also, the factors of "d" in the above three bounds may be replaced by $|\boldsymbol{L}|$, which is potentially much smaller.)

Remark 3.23. Given this corollary, it would be natural to fix

$$\eta = \sqrt{d/r}\cdot\tilde{\epsilon} \tag{89}$$

so as to balance the η term and the $(d/r)(\tilde{\epsilon}^2/\eta)$ term, making both contribute $O(\sqrt{d/r}\cdot\tilde{\epsilon})$. This swamps the $\tilde{\epsilon}\log(1/\tilde{\epsilon})$ term in Inequality (88) (up to a log factor). Thus with this choice of η we get the bound $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \leq \widetilde{O}(\sqrt{d/r}\cdot\tilde{\epsilon})$, and if we want this to equal some "ϵ_{final}" then we need to choose

$$\tilde{\epsilon} = \widetilde{\Theta}(\sqrt{r/d}\cdot\epsilon_{\text{final}}), \tag{90}$$

thereby making the final copy complexity $\widetilde{O}(rf/\tilde{\epsilon}) = \widetilde{O}(\sqrt{rd}\cdot f/\epsilon_{\text{final}})$, as stated in Theorem 1.6. Note that with these choices, the smallest eigenvalue of $\widehat{\boldsymbol{\rho}}$ will be $\widetilde{\Omega}(\epsilon_{\text{final}}/d)$, as expected.

The reason we do not simply directly fix $\eta = \sqrt{d/r}\cdot\tilde{\epsilon}$ in the proof of Corollary 3.22 is that in our later application to quantum zero mutual information testing (Section 4.3), it will be important to allow for a tradeoff between the off-diagonal χ^2-error, the on-diagonal χ^2-error, and the minimum eigenvalue of $\widehat{\boldsymbol{\rho}}$.

Proof of Corollary 3.22. Recall that in Theorem 3.18, we have

$$\epsilon = \tilde{\epsilon}\log(1/\tilde{\epsilon}); \tag{91}$$

we will use this shorthand throughout the present proof. We start by employing Inequality (36):

$$\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \leq \mathrm{D}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) + \widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) = \mathrm{D}^{\text{off}}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) + \mathrm{D}^{\text{on}}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) + \widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}). \tag{92}$$

To bound the off-diagonal contribution, we use: (i) each diagonal entry of $\widehat{\boldsymbol{\rho}}[\boldsymbol{L}]$ is at least $\eta/\boldsymbol{L}$; (ii) the off-diagonal Frobenius-squared of $\rho - \widehat{\boldsymbol{\rho}}$ is the same as that of $\rho - \widetilde{\boldsymbol{\rho}}$ (since $\widehat{\boldsymbol{\rho}}, \widetilde{\boldsymbol{\rho}}$ are both diagonal), which is bounded by $O(\tilde{\epsilon}^2)/(r\ln(1/\delta))$ from Item (c). Combining these facts yields

$$\mathrm{D}^{\text{off}}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) \leq O((|\boldsymbol{L}|/r)(\tilde{\epsilon}^2/\eta)). \tag{93}$$

As for the on-diagonal contribution, writing p for the diagonal entries of $\rho[\boldsymbol{L}]$, and $\widehat{\boldsymbol{q}} = (\eta/\boldsymbol{L})\mathbb{1}_{\boldsymbol{L}} + (1-\eta)\boldsymbol{q}$ for those of $\widehat{\boldsymbol{\rho}}$,

$$\mathrm{D}^{\text{on}}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) = \sum_{i\in\boldsymbol{L}} \frac{(p_i - \widehat{\boldsymbol{q}}_i)^2}{\widehat{\boldsymbol{q}}_i} \leq O(1)\sum_{i\in\boldsymbol{L}} \frac{(p_i - \boldsymbol{q}_i)^2}{\widehat{\boldsymbol{q}}_i} + O(1)\sum_{i\in\boldsymbol{L}} \frac{(\eta/\boldsymbol{L})^2}{\widehat{\boldsymbol{q}}_i} + O(1)\sum_{i\in\boldsymbol{L}} \frac{eta^2\boldsymbol{q}_i^2}{\widehat{\boldsymbol{q}}_i}. \tag{94}$$

For the first two summands above we use $\widehat{\boldsymbol{q}}_i \geq \eta^2/|\boldsymbol{L}|$ in the denominator; for the third, $\widehat{\boldsymbol{q}}_i \geq (1-\eta)\boldsymbol{q}_i \geq \frac{1}{2}\boldsymbol{q}_i$. Thus

$$\mathrm{D}^{\text{on}}_{\chi^2}(\rho[\boldsymbol{L}] \parallel \widehat{\boldsymbol{\rho}}[\boldsymbol{L}]) \leq O(|\boldsymbol{L}|/\eta)\|p[\boldsymbol{L}] - \boldsymbol{q}[\boldsymbol{L}]\|_2^2 + O(\eta) + O(\eta^2)\sum_{i\in\boldsymbol{L}} \widehat{\boldsymbol{q}}_i \leq O((|\boldsymbol{L}|/r)(\tilde{\eta}^2/\eta) + \eta), \tag{95}$$

where we used $\|p[\boldsymbol{L}] - \boldsymbol{q}[\boldsymbol{L}]\|_2^2 \leq \|\rho[\boldsymbol{L}] - \widetilde{\boldsymbol{\rho}}[\boldsymbol{L}]\|_{\mathrm{F}}^2$ and then Item (c) again.

In light of Inequalities (93) and (95), it suffices to show that $\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) = \widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel (1-\eta)\widetilde{\boldsymbol{\rho}}) \leq O(\eta+\epsilon)$, with the off-diagonal contribution being just $O(\epsilon)$. This off-diagonal contribution differs from that of $\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widetilde{\boldsymbol{\rho}})$ by a factor of at most $1/(1-\eta) = O(1)$, so we may indeed bound it by $O(\epsilon)$ from Item (d). Finally, the on-diagonal contribution to $\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}})$ is

$$\sum_{i\notin\boldsymbol{L}} \frac{(\rho_{ii} - (1-\eta)\boldsymbol{q}_i)^2}{(1-\eta)\boldsymbol{q}_i} \leq O(1)\sum_{i\notin\boldsymbol{L}} \frac{(\rho_{ii} - \boldsymbol{q}_i)^2}{\boldsymbol{q}_i} + O(1)\sum_{i\notin\boldsymbol{L}} \eta^2\boldsymbol{q}_i \leq O(\widehat{\mathrm{D}}^{-\boldsymbol{L}}_{\chi^2}(\rho \parallel \widetilde{\boldsymbol{\rho}})) + O(\eta^2), \tag{96}$$

and this is $O(\eta^2+\epsilon) \leq O(\eta+\epsilon)$, as required. □

Working out the parameters (just using Theorem 3.18), along the lines of Remark 3.23, we get:

Corollary 3.24. *A state estimation algorithm with Frobenius-squared rate $f = f(d,r) \gg \log d$ may be transformed (preserving the single-copy measurement property) into a state estimation algorithm with the following property:*

Given parameters ϵ, r (with $\epsilon \leq 1/2$), and M copies of a quantum state $\rho \in \mathbb{C}^{d\times d}$ of rank at most r,

$$M = \widetilde{O}\left(\frac{\sqrt{rd}\cdot f(d,r)}{\epsilon}\right) \tag{97}$$

suffices[6] *for the algorithm to output (with probability at least .99) the classical description of a quantum state* $\widehat{\boldsymbol{\rho}}$ *with* $\mathrm{D}_{\chi^2}(\rho \parallel \widehat{\boldsymbol{\rho}}) \leq \epsilon$.

In particular, by Theorem 1.3

$$M = \widetilde{O}\left(\frac{r^{.5}d^{1.5}}{\epsilon}\right) \text{ suffices using collective measurements.} \tag{98}$$

And, by Theorem 1.1

$$M = \widetilde{O}\left(\frac{r^{1.5}d^{1.5}}{\epsilon}\right) \text{ suffices using single-copy measurements.} \tag{99}$$

3.6 A simple Frobenius-squared estimator

Proposition 3.25. *There is estimation algorithm for quantum states, making single-copy measurements, achieving expected Frobenius-squared error* $O(d^2/m)$.

Proof. It suffices to achieve Frobenius-squared error d/m using $O(dm)$ copies. Assume for simplicity d is even. (We leave the odd d case to the reader.) Then there is a simple way [47] to construct a partition $\mathcal{P}$ of the edges of the complete graph on d vertices into $d-1$ matchings. Fix a particular matching $M \in \mathcal{P}$, and associate to it the POVM with elements

$$(X_{ij}^{\pm})_{\{i,j\}\in M}, \quad \text{where} \quad X_{ij}^{\pm} = \frac{1}{2}(|i\rangle\langle i| + |j\rangle\langle j|) \pm \frac{1}{2}(|i\rangle\langle j| + |j\rangle\langle i|). \tag{100}$$

When we measure ρ with this POVM, we obtain outcome $(\{i,j\},\pm)$ with probability

$$p_{ij} \pm r_{ij}, \quad p_{ij} \coloneqq \mathrm{avg}\{\rho_{ii},\rho_{jj}\}, \quad r_{ij} \coloneqq \mathrm{Re}\,\rho_{ij}. \tag{101}$$

If we similarly define a POVM $(Y_{ij}^{\pm})$ but with a factor of $\mathrm{i} = \sqrt{-1}$ in the off-diagonal elements of Equation (100), we will similarly get outcomes with probabilities $p_{ij} \pm s_{ij}$, where $s_{ij} \coloneqq \mathrm{Im}\,\rho_{ij}$. We focus on analyzing Equation (101), as the imaginary-part analysis will be identical.

Suppose we now measure this POVM m times and form the random variables $\widehat{\boldsymbol{r}}_{ij}$ (for $\{i,j\} \in M$), where

$$\widehat{\boldsymbol{r}}_{ij} = \frac{\boldsymbol{f}_{ij}^{+} - \boldsymbol{f}_{ij}^{-}}{2}, \quad \text{with} \quad \boldsymbol{f}_{ij}^{\pm} = \text{fraction of outcomes that are } (\{i,j\},\pm). \tag{102}$$

Then $\mathbf{E}[\widehat{\boldsymbol{r}}_{ij}] = r_{ij}$, and

$$\mathbf{E}[(r_{ij} - \widehat{\boldsymbol{r}}_{ij})^2] = \mathbf{Var}[\widehat{\boldsymbol{r}}_{ij}] \leq \frac{1}{2}\mathbf{Var}[\boldsymbol{f}_{ij}^{+}] + \frac{1}{2}\mathbf{Var}[\boldsymbol{f}_{ij}^{-}] \leq \frac{p_{ij}+r_{ij}}{2m} + \frac{p_{ij}-r_{ij}}{2m} = p_{ij}/m. \tag{103}$$

Repeating the analysis for the imaginary parts, we use $2m$ copies of ρ to get estimates for all ρ_{ij}, $\{i,j\} \in M$, achieving total expected squared-error

$$\sum_{\{i,j\}\in M} 2p_{ij}/m = \sum_{i\in[d]} \rho_{ii}/m = 1/m. \tag{104}$$

Repeating this for all $M \in \mathcal{P}$ uses $O(dm)$ copies of ρ and gives estimates for all off-diagonal elements of ρ, with total expected squared-error $(d-1)/m$. Finally, we can use standard basis measurements to estimate the diagonal elements of ρ, using Proposition 2.14: m more copies of ρ suffice to achieve total expected squared-error $1/m$. □

[6] The hidden polylog terms are at most $\log^2(d/\epsilon)\log\log(d/\epsilon)$, but may be optimized further in special cases [41]. In case $r = d$, one may take $M = O(\frac{df(d)}{\epsilon}\log^2(1/\epsilon)\log\log(1/\epsilon))$, so the polylog terms have no dependence on d. In case $r = O(1)$, if $\epsilon \geq \exp\left(-\Omega(\sqrt{d})\right)$, one may take $M = O(\frac{\sqrt{d}f(d)}{\epsilon}\log(d/\epsilon)\log\log(d/\epsilon))$.

4 Testing zero mutual information

We now move on to showing the main application of our χ^2 tomography algorithm: testing zero quantum mutual information. We will explain below in Section 4.3 how our χ^2 tomography algorithm is crucial to achieving this result. But first, we introduce and analyze a variant of the quantum mutual information that features in our analysis.

4.1 Mutual information versus its Hellinger variant

The goal of this subsection is to prove the below theorem, showing that the standard quantum mutual information is not much larger than the "Hellinger mutual information":

Theorem 4.1. *Let $\rho = \rho_{AB}$ be a bipartite quantum state on $A \otimes B$, where $A \cong B \cong \mathbb{C}^d$. Writing $\mathrm{D}^2_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B) = \eta$, it holds that $I(A:B)_\rho \leq \eta \cdot O(\log(d/\eta))$.*

We also observe that by restricting ρ_{AB} to be diagonal, we immediately obtain the analogous theorem concerning classical mutual information. We remark that proving this classical version directly is no easier than proving the quantum version.

Corollary 4.2. *Let $p = p_{AB}$ be bipartite classical state on $A \times B$, where $|A| = |B| = d$. Writing $\mathrm{d}^2_{\mathrm{H}}(p, p_A \times p_B) = \eta$, it holds that $I(A:B)_p \leq \eta \cdot O(\log(d/\eta))$.*

We first state a bound on the continuity of mutual information in terms of the trace distance and the subsystem dimension. A bound of the following form can be proven a number of ways, for example by appealing to the Petz–Fannes–Audenaert [4] and Alicki–Fannes [2] inequalities; see [18, Appendix F]. The bound we use is an immediate corollary of [40, Prop. 1] which gives small explicit constants.

Lemma 4.3 (Continuity of quantum mutual information)**.** *For two density operator ρ, σ on $A \otimes B$ with $A \cong B \cong \mathbb{C}^d$ and $\frac{1}{2}\|\rho - \sigma\|_1 = \epsilon$, we have*

$$|I(A:B)_\rho - I(A:B)_\sigma| \leq 2\epsilon \log \frac{4d}{\epsilon}\,. \tag{105}$$

Proof. The bound from [40, Prop. 1] for the quantum conditional mutual information applies immediately with a trivial conditioning system and with the bound $\epsilon \leq 1$ to get the constant 4. □

We also must bound how much $\mathrm{D}^2_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B)$ changes relative to a depolarization of ρ.

Lemma 4.4. *Given a density operator ρ on $A \otimes B$ with $A \cong B \cong \mathbb{C}^d$ and the depolarization $\sigma = \Delta_\epsilon(\sigma) = (1-\epsilon)\rho + \frac{\epsilon}{d^2}\mathbb{1}$, the squared Hellinger distance obeys*

$$\mathrm{D}^2_{\mathrm{H}}(\sigma, \sigma_A \otimes \sigma_B) \leq C\sqrt{\epsilon} + \mathrm{D}^2_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B)\,, \tag{106}$$

where the constant can be chosen as $C = 4 + 4\sqrt{2}$.

Proof. Since $\mathrm{D}_{\mathrm{H}}(\rho, \sigma)$ is a metric, we have

$$\mathrm{D}_{\mathrm{H}}(\sigma, \sigma_A \otimes \sigma_B) \leq \mathrm{D}_{\mathrm{H}}(\sigma, \rho) + \mathrm{D}_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B) + \mathrm{D}_{\mathrm{H}}(\rho_A \otimes \rho_B, \sigma_A \otimes \sigma_B)\,. \tag{107}$$

By Proposition 2.31 we have $\mathrm{D}_{\mathrm{H}}(\rho, \sigma) \leq \sqrt{2\mathrm{D}_{\mathrm{tr}}(\rho, \sigma)}$ and we have $\mathrm{D}_{\mathrm{tr}}(\rho, \sigma) \leq \epsilon$ by the triangle inequality. By the triangle inequality and quantum data processing inequality (monotonicity of trace distance under CPTP maps), we also have $\mathrm{D}_{\mathrm{tr}}(\rho_A \otimes \rho_B, \sigma_A \otimes \sigma_B) \leq \mathrm{D}_{\mathrm{tr}}(\rho_A, \sigma_A) + \mathrm{D}_{\mathrm{tr}}(\rho_B, \sigma_B) \leq 2\epsilon$. Plugging this in above and rearranging, we have

$$\mathrm{D}_{\mathrm{H}}(\sigma, \sigma_A \otimes \sigma_B) - \mathrm{D}_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B) \leq (2 + \sqrt{2})\sqrt{\epsilon}\,. \tag{108}$$

For any two states ρ, σ, $\mathrm{D}_{\mathrm{H}}(\rho, \sigma) \leq \sqrt{2}$, so we can multiply both sides by $\mathrm{D}_{\mathrm{H}}(\sigma, \sigma_A \otimes \sigma_B) + \mathrm{D}_{\mathrm{H}}(\rho, \rho_A \otimes \rho_B) \leq 2\sqrt{2}$ and the bound follows. □

Now we can prove Theorem 4.1.

Proof of Theorem 4.1. We begin by smoothing the state ρ with a depolarizing channel to obtain

$$\sigma = (1-\epsilon)\rho + \frac{\epsilon}{d^2}\mathbb{1}\,. \tag{109}$$

By the triangle inequality, we have $\frac{1}{2}\|\rho-\sigma\|_1 = \frac{\epsilon}{2}\|\rho - \mathbb{1}/d^2\|_1 \leq \epsilon$. The change in the mutual information by passing from $\rho \to \sigma$ is bounded using Lemma 4.3,

$$I(A:B)_\rho \leq I(A:B)_\sigma + 2\epsilon \log \frac{4d}{\epsilon}\,. \tag{110}$$

Note that $I(A:B)_\rho \geq I(A:B)_\sigma$ by the quantum data processing inequality for relative entropy (monotonicity of mutual information under local CPTP maps).

By Theorem 2.32, we have

$$I(A:B)_\sigma \leq (2 + \mathrm{D}_\infty^{\mathrm{Rén}}(\sigma \parallel \sigma_A \otimes \sigma_B)) \cdot \mathrm{D}_{\mathrm{H}}^2(\sigma, \sigma_A \otimes \sigma_B)\,. \tag{111}$$

Since $\sigma - \frac{\epsilon}{d^2}\mathbb{1} \geq 0$, the positivity of the partial trace map shows that the reduced states σ_A and σ_B satisfy

$$\sigma_A \geq \frac{\epsilon}{d}\mathbb{1}\,, \text{ and } \quad \sigma_B \geq \frac{\epsilon}{d}\mathbb{1}\,. \tag{112}$$

Therefore the Rényi entropy term is bounded using Fact 2.29 as

$$\mathrm{D}_\infty^{\mathrm{Rén}}(\sigma \parallel \sigma_A \otimes \sigma_B)) \leq \log \|\sigma_A^{-1} \otimes \sigma_B^{-1}\| \leq \log \frac{d^2}{\epsilon^2}\,. \tag{113}$$

Using Lemma 4.4, the $\mathrm{D}_{\mathrm{H}}^2(\sigma, \sigma_A \otimes \sigma_B)$ term is bounded by

$$\mathrm{D}_{\mathrm{H}}^2(\sigma, \sigma_A \otimes \sigma_B) \leq C\sqrt{\epsilon} + \mathrm{D}_{\mathrm{H}}^2(\rho, \rho_A \otimes \rho_B) = C\sqrt{\epsilon} + \eta\,. \tag{114}$$

Putting this all together, we have that

$$I(A:B)_\rho \leq \Big(2 + \log \frac{d^2}{\epsilon^2}\Big) \cdot \big(C\sqrt{\epsilon} + \eta\big) + 2\epsilon \log \frac{4d}{\epsilon} \leq (\eta + \sqrt{\epsilon})\, O\big(\log(d/\epsilon)\big)\,. \tag{115}$$

Choosing $\epsilon = O(\eta^2)$ completes the proof. □

4.2 Testing zero classical mutual information

Before moving to the trickier quantum case, we warm up by showing an efficient tester for classical mutual information, establishing Theorem 1.18. First we give a short proof of the following (which appears explicitly as [1, Lem. 7]):

Lemma 4.5. *Given $n = O(d/\epsilon)$ samples from two distributions q, q' on $[d]$, one can output a hypothesis $\widehat{\boldsymbol{q}} \times \widehat{\boldsymbol{q}}'$ that (with probability at least .99) satisfies $\mathrm{d}_{\chi^2}(q \times q' \parallel \widehat{\boldsymbol{q}} \times \widehat{\boldsymbol{q}}') \leq \epsilon$.*

Proof. It suffices to separately learn each of q, q' to χ^2-accuracy $\epsilon/3$ and high confidence (which can be done applying Proposition 2.16 and Markov's inequality), and then apply the below Proposition 4.6. □

Proposition 4.6. *Let $\mathrm{d}_{\chi^2}(p \parallel q) = \epsilon_1$, $\mathrm{d}_{\chi^2}(p' \parallel q') = \epsilon_2$. Then we have the near-additivity formula*

$$\mathrm{d}_{\chi^2}(p \otimes p' \parallel q \otimes q') = (1+\epsilon_1)(1+\epsilon_2) - 1 = \epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2. \tag{116}$$

Proof. This follows essentially immediate from the second formula in Definition 2.9. □

Now to prove Theorem 1.18, suppose we are given access to a bipartite probability distribution $p = p_{AB}$ on $[d] \times [d]$ and we are promised that either $I(A:B)_p = 0$ or $I(A:B)_p \geq \epsilon$. In the former case we have

$$I(A:B)_p = 0 \quad \Longrightarrow \quad p = p_A \times p_B \quad \Longrightarrow \quad \mathrm{d}_{\chi^2}(p \parallel \widehat{\boldsymbol{p}}_A \times \widehat{\boldsymbol{p}}_B) \leq c\epsilon/\log(d/\epsilon) \tag{117}$$

with high probability if we estimate the two marginals using $O((d/\epsilon)\cdot\log(d/\epsilon))$ samples as in Lemma 4.5. (Here $c>0$ may be any small constant.) On the other hand, in case $I(A:B)_p \geq \epsilon$, we get $\mathrm{d}_{\mathrm{H}}^2(p, p_A\times p_B)\geq \Omega(\epsilon/\log(d/\epsilon))$ from Corollary 4.2; ensuring that the constant c in Equation (117) is small enough, this implies

$$\mathrm{d}_{\mathrm{H}}^2(p, \widehat{\boldsymbol{p}}_A\times\widehat{\boldsymbol{p}}_B)\geq\Omega(\epsilon/\log(d/\epsilon)) \tag{118}$$

also. Now again ensuring c is small enough, our classical mutual information tester Theorem 1.18 follows by using the below "χ^2-vs.-H^2 identity tester" of Daskalakis–Kamath–Wright with the "known" distribution being $\widehat{\boldsymbol{p}}_A\times\widehat{\boldsymbol{p}}_B$. This distinguishes our two cases (with high probability) using $O(\sqrt{d\times d}/\epsilon')$ samples, $\epsilon'=\Theta(\epsilon/\log(d/\epsilon))$; in other words, with $O((d/\epsilon)\log(d/\epsilon))$ samples.

Theorem 4.7. *([15, Thm. 1].) For any "known" distribution q on $[d]$, there is a testing algorithm with the following guarantee: Given $0<\epsilon\leq 1/2$, as well as $n=O(\sqrt{d}/\epsilon)$ samples from an unknown distribution p on $[d]$, if $\mathrm{d}_{\chi^2}(p\parallel q)\leq\epsilon/2$ then the test accepts with probability at least .99, and if $\mathrm{d}_{\mathrm{H}}^2(p,q)\geq\epsilon$ then the test rejects with probability at least .99.*

4.3 Testing zero quantum mutual information

To prove Theorem 1.14, we now endeavor to repeat the result from the previous setting in the quantum case. Naturally, it is crucially important that we are able to do quantum state tomography with respect to Bures χ^2-divergence. This lets use the following quantum "χ^2-vs.-H^2 identity tester" from [7] in place of Theorem 4.7.

Theorem 4.8. *([7, Thm. 1].) For any "known" quantum state $\sigma\in\mathbb{C}^{d\times d}$, there is a testing algorithm with the following guarantee: Given $0<\epsilon\leq 1/2$, as well as $n=O(d/\epsilon)$ copies of an unknown state $\rho\in\mathbb{C}^{d\times d}$, if $\mathrm{D}_{\chi^2}(\rho\parallel\sigma)\leq .49\epsilon$ then the test accepts with probability at least .99, and if $\mathrm{D}_{\mathrm{H}}^2(\rho,\sigma)\geq 2\epsilon$ then the test rejects with probability at least .99.*

We can relate quantum mutual information and its Hellinger version using Theorem 4.1 in place of Corollary 4.2. It would seem then that we could establish our quantum zero mutual information tester Theorem 1.14 in exactly the same way we did its classical analogue, using $\widetilde{O}(d^2/\epsilon)$ copies of a bipartite $d\times d$-dimensional state ρ. Unfortunately, there is a missing piece: a quantum analogue of Lemma 4.5 with $O(d^2/\epsilon)$ copy complexity. We prove a slightly worse variant, which leads to our main testing Theorem 1.14:

Theorem 4.9. *There is a tomography algorithm that, given parameter $0<\epsilon\leq 1/2$ and*

$$n=\max\{\widetilde{O}(rd^{1.5}/\epsilon),\ \widetilde{O}(r^{.5}d^{1.75}/\epsilon)\} \tag{119}$$

copies of unknown d-dimensional quantum states ξ,ρ of rank at most r, outputs diagonal states $\boldsymbol{\sigma},\boldsymbol{\tau}$ such that (with probability at least .9),

$$\mathrm{D}_{\chi^2}(\xi\otimes\rho\parallel\boldsymbol{\sigma}'\otimes\boldsymbol{\tau}')\leq\epsilon. \tag{120}$$

Proof. The strategy is to apply our central estimation algorithm in the form of Corollary 3.22 to both ξ,ρ (with the Frobenius-learner from Theorem 1.3 with $f(d,r)=O(d)$). We use the parameter choices

$$\eta=\epsilon,\qquad \tilde{\epsilon}=\epsilon\cdot\min\{1/d^{.5}, r^{.5}/d^{.75}\}. \tag{121}$$

This leads to the claimed copy complexity from Equation (119), and yields (with probability at least .9) estimates $\boldsymbol{\sigma},\boldsymbol{\tau}$ satisfying

$$\mathrm{D}_{\chi^2}^{\mathrm{off}}(\xi\parallel\boldsymbol{\sigma}),\ \mathrm{D}_{\chi^2}^{\mathrm{off}}(\rho\parallel\boldsymbol{\tau})\leq\widetilde{O}(\epsilon\cdot(1/d^{.5}+\min\{1/r, r^{.5}/d^{.75}\}))=\widetilde{O}(\epsilon/\sqrt{d}), \tag{122}$$

$$\mathrm{D}_{\chi^2}^{\mathrm{on}}(\xi\parallel\boldsymbol{\sigma}),\ \mathrm{D}_{\chi^2}^{\mathrm{off}}(\rho\parallel\boldsymbol{\tau})\leq\widetilde{O}(\epsilon), \tag{123}$$

with $\boldsymbol{\sigma},\boldsymbol{\tau}$ having minimum eigenvalue at least ϵ/d. We claim that for any fixed outcomes $\boldsymbol{\sigma}=\sigma'$ and $\boldsymbol{\tau}'=\tau'$ satisfying the above, it holds that

$$\mathrm{D}_{\chi^2}(\xi\otimes\rho\parallel\sigma'\otimes\tau')\leq\widetilde{O}(\epsilon). \tag{124}$$

This is sufficient to complete the proof.

To establish Inequality (124), let us write $\sigma' = \mathrm{diag}(s')$ and $\tau' = \mathrm{diag}(t')$, with

$$s'_a \geq \epsilon/d, \quad t'_j \geq \epsilon/d \qquad \text{for all } a, j \in [d]. \tag{125}$$

We will break up $\mathrm{D}_{\chi^2}(\xi \otimes \rho \parallel \sigma' \otimes \tau')$ into three parts: on-on-diagonal, on-off-diagonal, and off-off-diagonal:

$$\underbrace{\sum_{a,i=1}^{d} \tfrac{1}{s'_a t'_i} (\xi_{aa}\rho_{ii} - s'_a t'_i)^2}_{\text{(ON-ON)}} + \underbrace{\left(\sum_{a=1}^{d} \sum_{i \neq j} \tfrac{2}{s'_a t'_i + s'_a t'_j} |\xi_{aa}\rho_{ij}|^2 + \sum_{a \neq b} \sum_{i=1}^{d} \tfrac{2}{s'_a t'_i + s'_b t'_i} |\xi_{ab}\rho_{ii}|^2 \right)}_{\text{(ON-OFF)}} + \underbrace{\sum_{a \neq b} \sum_{i \neq j} \tfrac{2}{s'_a t'_i + s'_b t'_j} |\xi_{ab}\rho_{ij}|^2}_{\text{(OFF-OFF)}}. \tag{126}$$

First, using Proposition 4.6,

$$\text{(ON-ON)} = \mathrm{d}_{\chi^2}(\mathrm{diag}(\xi) \otimes \mathrm{diag}(\rho) \parallel s' \otimes t') = (1 + \mathrm{d}_{\chi^2}(\mathrm{diag}(\xi) \parallel s'))(1 + \mathrm{d}_{\chi^2}(\mathrm{diag}(\rho) \parallel t')) - 1. \tag{127}$$

But $\mathrm{d}_{\chi^2}(\mathrm{diag}(\xi) \parallel s') \leq \mathrm{D}^{\mathrm{on}}_{\chi^2}(\xi \parallel \sigma') \leq \widetilde{O}(\epsilon)$ by Inequality (122), and similarly for $\mathrm{d}_{\chi^2}(\mathrm{diag}(\rho) \parallel t')$, so we conclude from Inequality (127) that (ON-ON) $\leq (1 + \widetilde{O}(\epsilon))(1 + \widetilde{O}(\epsilon)) - 1 = \widetilde{O}(\epsilon)$, as needed for Inequality (124).

Moving to (ON-OFF), the first term in it factorizes to

$$\left(\sum_{a=1}^{d} \tfrac{1}{s'_a} \xi_{aa}^2 \right) \left(\sum_{i \neq j} \tfrac{2}{t'_i + t'_j} |\rho_{ij}|^2 \right) \tag{128}$$

The first factor above is precisely

$$1 + \mathrm{d}_{\chi^2}(\mathrm{diag}(\xi) \parallel s') \leq 1 + \mathrm{D}^{\mathrm{on}}_{\chi^2}(\xi \parallel \sigma') \leq 1 + \widetilde{O}(\epsilon) \leq O(1). \tag{129}$$

The second factor in Equation (128) is

$$\sum_{i \neq j} \tfrac{2}{t'_i + t'_j} |\rho_{ij}|^2 = \mathrm{D}^{\mathrm{off}}_{\chi^2}(\rho \parallel \tau') \leq \mathrm{D}_{\chi^2}(\rho \parallel \tau) t' \leq \widetilde{O}(\epsilon). \tag{130}$$

Thus Equation (128), and indeed both terms in (ON-OFF), can be bounded by $\widetilde{O}(\epsilon)$, as needed for Inequality (124).

It remains to bound (OFF-OFF) by $\widetilde{O}(\epsilon)$. By the AM-GM inequality,

$$\frac{s'_a t'_i + s'_b t'_j}{2} \geq \sqrt{s'_a t'_i s'_b t'_j} = \sqrt{s'_a s'_b} \sqrt{t'_i t'_j}. \tag{131}$$

Of course there is no reverse AM-GM inequality, but we at least have

$$\sqrt{xy} \geq \min(\sqrt{x/y}, \sqrt{y/x}) \cdot \frac{x+y}{2} \qquad \forall x, y > 0. \tag{132}$$

When (x, y) is (s'_a, s'_b) or (t'_i, t'_j), we have $\min(\sqrt{x/y}, \sqrt{y/x}) \geq \sqrt{\epsilon/d}$ (from Inequality (125)), and hence

$$\frac{s'_a t'_i + s'_b t'_j}{2} \geq (\epsilon/d) \cdot \frac{s'_a + s'_b}{2} \cdot \frac{t'_i + t'_j}{2} \tag{133}$$

Putting this into the definition of (OFF-OFF) yields

$$\text{(OFF-OFF)} \leq (d/\epsilon) \cdot \sum_{a \neq b} \sum_{i \neq j} \tfrac{2}{s'_a + s'_b} \cdot \tfrac{2}{t'_i + t'_j} \cdot |\xi_{ab}\rho_{ij}|^2 = (d/\epsilon) \left(\sum_{a \neq b} \tfrac{2}{s'_i + s'_j} |\xi_{ab}|^2 \right) \left(\sum_{i \neq j} \tfrac{2}{t'_i + t'_j} |\rho_{ij}|^2 \right). \tag{134}$$

The last factor here is bounded by $\mathrm{D}^{\mathrm{off}}_{\chi^2}(\rho \parallel \tau')$ in Equation (130), and the similar factor with s' and ξ is similarly bounded. Hence using Inequality (122), we indeed get

$$\text{(OFF-OFF)} \leq O(d/\epsilon) \cdot \widetilde{O}(\epsilon/\sqrt{d}) \cdot \widetilde{O}(\epsilon/\sqrt{d}) = \widetilde{O}(\epsilon), \tag{135}$$

completing the proof. □

5 Open Problems

One obvious and by now longstanding open question related to our work is learning in infidelity to precision ϵ with $O(rd/\epsilon)$ samples, without any logarithms. This would settle the sample complexity of tomography with infidelity loss up to constant factors. In light of our work, perhaps we could even ask for more: Given our result that learning in quantum relative entropy is possible with $\widetilde{O}(rd/\epsilon)$ samples, might a similar no-logarithm bound hold here as well?

Our algorithm uses only single-copy measurements, but even these are challenging on present-day quantum computers. A stronger assumption on measurements is to restrict to product measurements, meaning that all POVM elements factorize into tensor products over subsystems. We believe this measurement model will require strictly greater sample complexity for learning in χ^2-divergence and for quantum mutual information testing than the single-copy case analyzed here.

Regarding quantum mutual information testing, note that in the classical case we could learn product states to χ^2-divergence well enough that the entire testing complexity was dominated by the χ^2-vs.-Hellinger identity tester. Unfortunately, in the quantum case we couldn't quite match this. Might it be possible to reduce the complexity of testing zero quantum mutual information down to to $\widetilde{O}(d^2/\epsilon)$?

For learning in χ^2-divergence, it would be interesting to show that $\widetilde{\Omega}(\sqrt{r}d^{1.5}/\epsilon)$ is the right lower bound; currently, we have nothing better than the infidelity-tomography lower bound of $\widetilde{\Omega}(rd/\epsilon)$. As explained in Remark 3.17, though, it seems like reducing the upper bound could be difficult even for the case $r = 1$.

Although the *Bures* χ^2-divergence is usually the largest of the "big four" quantities considered in this paper, there are other quantum generalizations of χ^2-divergence in the literature that are larger still than Bures χ^2-divergence (see, e.g., [37, 42]). An example is the so-called "standard" quantum χ^2-divergence, in which the the arithmetic mean reciprocal-prefactor in Equation (33) is replaced by a geometric mean. Similarly, there are also multiple generalizations of the quantum relative entropy besides the "Umegaki" quantum relative entropy $\mathrm{S}(\cdot \,\|\, \cdot)$ studied herein. As explained above, the main reason for us to consider learning with respect to *Bures* χ^2-divergence (as opposed to other metrics) is that it seems necessary for some applications; for example, our quantum mutual information testing problem. It is an interesting open question to study state tomography with respect to other generalizations of relative entropy and χ^2-divergence, and in particular to decide if this is possible while still having $\widetilde{O}(1/\epsilon)$ scaling.

More generally, a very interesting direction is to investigate for which quantum learning and testing tasks we can get away with $\widetilde{O}(1/\epsilon)$ samples, and for which we require (say) $\widetilde{\Omega}(1/\epsilon^2)$ samples.

References

[1] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[2] Robert Alicki and Mark Fannes. Continuity of quantum conditional information. *Journal of Physics A: Mathematical and General*, 37(5):L55, 2004.

[3] George Androulakis and Tiju John. Quantum f-divergences via Naussbaum–Szkoła distributions with applications to Petz–Rényi and von Neumann relative entropy. Technical Report 2203.01964, arXiv, 2022.

[4] Koenraad Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127, 2007.

[5] Koenraad Audenaert and Jens Eisert. Continuity bounds on the quantum relative entropy–II. *Journal of Mathematical Physics*, 52(11):112201, 2011.

[6] Koenraad Audenaert, Michael Nussbaum, Arleta Szkoła, and Frank Verstraete. Asymptotic error rates in quantum hypothesis testing. *Communications in Mathematical Physics*, 279(1):251–283, feb 2008.

[7] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. ACM, June 2019.

[8] Emilio Bagan, Manuel Ballester, Richard Gill, Alex Monràs, and Ramon Muñoz-Tapia. Optimal full estimation of qubit mixed states. *Physical Review A*, 73(3):032301, 2006.

[9] Mario Berta, Omar Fawzi, and Marco Tomamichel. On variational expressions for quantum relative entropies. *Letters in Mathematical Physics*, 107(12):2239–2265, 2017.

[10] Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by Chow–Liu. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. ACM, June 2021.

[11] Robin Blume-Kohout and Patrick Hayden. Accurate quantum state estimation via "keeping the experimentalist honest", 2006.

[12] Samuel Braunstein and Carlton Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439–3443, 1994.

[13] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. When does adaptivity help for quantum state learning? Technical report, arXiv:2206.00220, 2022.

[14] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. When does adaptivity help for quantum state learning?, 2023. New version of [CHLLS22] in preparation.

[15] Constantinos Daskalakis, Gautam Kamath, and John Wright. *Which Distribution Distances are Sublinearly Testable?*, pages 2747–2764. Society for Industrial and Applied Mathematics, 2018.

[16] Nilanjana Datta. Min- and max-relative entropies and a new entanglement monotone. *IEEE Transactions on Information Theory*, 55(6):2816–2826, June 2009.

[17] Christopher Ferrie and Robin Blume-Kohout. Minimax quantum tomography: the ultimate bounds on accuracy. Technical Report 1503.03100, arXiv, 2015.

[18] Steven T. Flammia, Jeongwan Haah, Michael Kastoryano, and Isaac Kim. Limits on the storage of quantum information in a volume of space. *Quantum*, 1:4, April 2017.

[19] Christopher Fuchs and Carlton Caves. Mathematical techniques for quantum communication theory. *Open Systems & Information Dynamics*, 3(3):345–356, October 1995.

[20] Alison Gibbs and Francis Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419, December 2002.

[21] Jeongwan Haah, Aram Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, pages 1–1, 2017.

[22] Jeongwan Haah, Robin Kothari, Ryan O'Donnell, and Ewin Tang. Query-optimal estimation of unitary channels in diamond distance. Technical Report 2302.14066, arXiv, 2023.

[23] Aram Harrow and Ashley Montanaro. Testing product states, quantum Merlin–Arthur games and tensor optimization. *Journal of the ACM*, 60(1):1–43, 2013.

[24] Masahito Hayashi. Asymptotic estimation theory for a finite-dimensional pure state model. *Journal of Physics A: Mathematical and General*, 31(20):4633–4655, may 1998.

[25] Carl Helstrom. *Quantum detection and estimation theory*, volume 123 of *Mathematics in science and engineering*. Academic Press, New York, 1 edition, 1976.

[26] Fumio Hiai and Milán Mosonyi. Different quantum f-divergences and the reversibility of quantum operations. *Rev. Math. Phys.*, 29(7):1750023, 80, 2017.

[27] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[28] Martin Müller-Lennert, Frédéric Dupuis, Oleg Szehr, Serge Fehr, and Marco Tomamichel. On quantum Rényi entropies: A new generalization and some properties. *Journal of Mathematical Physics*, 54(12):122203, December 2013.

[29] Michael Nielsen and Isaac Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th Anniversary edition, 2010.

[30] Michael Nussbaum and Arleta Szkoła. The Chernoff lower bound for symmetric quantum hypothesis testing. *Annals of Statistics*, 37(2), apr 2009.

[31] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 899–912, New York, NY, USA, 2016. Association for Computing Machinery.

[32] Ryan O'Donnell and John Wright. Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 962–974, New York, NY, USA, 2017. Association for Computing Machinery.

[33] Luciana Pereira, Leonardo Zambrano, Jean Cortés-Vega, Sebastián Niklitschek, and Aldo Delgado. Adaptive quantum tomography in high dimensions. *Physical Review A*, 98(1):012339, July 2018.

[34] Luciano Pereira, Juan García-Ripoll, and Tomá Ramos. Parallel tomography of quantum non-demolition measurements in multi-qubit devices. *npj Quantum Information*, 9(1):22, 2023.

[35] Yuval Peres. Finite-sample deviation bound of empirical distribution from true distribution. MathOverflow, 2019. `https://mathoverflow.net/q/331405`.

[36] Dénes Petz. Quasi-entropies for finite quantum systems. *Reports on Mathematical Physics*, 23(1):57–65, February 1986.

[37] Dénes Petz. Monotone metrics on matrix spaces. *Linear Algebra and its Applications*, 244:81–96, September 1996.

[38] Jaroslav Řeháček, Berthold-Georg Englert, and Dagomir Kaszlikowski. Minimal qubit tomography. *Physical Review A*, 70(5):052321, November 2004.

[39] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[40] M. E. Shirokov. Tight uniform continuity bounds for the quantum conditional mutual information, for the Holevo quantity, and for capacities of quantum channels. *Journal of Mathematical Physics*, 58(10):102202, October 2017.

[41] Ewin Tang, 2022. `https://twitter.com/ewintang/status/1569426816821248001?cxt=HHwWgoDS2f3X3McrAAAA`.

[42] Kristan Temme, Michael Kastoryano, Mary Beth Ruskai, Michael Wolf, and Frank Verstraete. The χ^2-divergence and mixing times of quantum Markov processes. *Journal of Mathematical Physics*, 51(12):122201, 2010.

[43] Kristan Temme and Frank Verstraete. Quantum chi-squared and goodness of fit testing. *Journal of Mathematical Physics*, 56(1):012202, January 2015.

[44] Marco Tomamichel. *Quantum Information Processing with Finite Resources*. Springer International Publishing, 2016.

[45] Marco Tomamichel and Jan Seyfried. Personal communication, 2024.

[46] Hisaharu Umegaki. Conditional expectation in an operator algebra. IV. Entropy and information. *Kodai Mathematical Journal*, 14(2), January 1962.

[47] Wikipedia. Graph factorization (complete graphs), 2023. https://en.wikipedia.org/wiki/Graph_factorization#Complete_graphs.

[48] Mark Wilde, Andreas Winter, and Dong Yang. Strong converse for the classical capacity of entanglement-breaking and Hadamard channels via a sandwiched Rényi relative entropy. *Communications in Mathematical Physics*, 331(2):593–622, jul 2014.

[49] Andreas Winter. Coding theorem and strong converse for quantum channels. *IEEE Transactions on Information Theory*, 45(7):2481–2485, 1999.

[50] Yihong Wu. Lecture notes: Information-theoretic methods for high-dimensional statistics. www.stat.yale.edu/~yw562/teaching/it-stats.pdf, 2020.

[51] Henry Yuen. An improved sample complexity lower bound for (fidelity) quantum state tomography. *Quantum*, 7:890, 2023.