Lower bounds for learning quantum states with single-copy measurements

Angus Lowe * Ashwin Nayak †

July 27, 2022

Abstract

We study the problems of quantum tomography and shadow tomography using measurements performed on individual, identical copies of an unknown d-dimensional state. We first revisit known lower bounds [HHJ+17] on quantum tomography with accuracy ϵ in trace distance, when the measurement choices are independent of previously observed outcomes, i.e., they are nonadaptive. We give a succinct proof of these results through the χ^2-divergence between suitable distributions. Unlike prior work, we do not require that the measurements be given by rank-one operators. This leads to stronger lower bounds when the learner uses measurements with a constant number of outcomes (e.g., two-outcome measurements). In particular, this rigorously establishes the optimality of the folklore "Pauli tomography" algorithm in terms of its sample complexity. We also derive novel bounds of $\Omega(r^2 d/\epsilon^2)$ and $\Omega(r^2 d^2/\epsilon^2)$ for learning rank r states using arbitrary and constant-outcome measurements, respectively, in the nonadaptive case.

In addition to the sample complexity, a resource of practical significance for learning quantum states is the number of unique measurement settings required (i.e., the number of different measurements used by an algorithm, each possibly with an arbitrary number of outcomes). Motivated by this consideration, we employ concentration of measure of χ^2-divergence of suitable distributions to extend our lower bounds to the case where the learner performs possibly adaptive measurements from a fixed set of $\exp(O(d))$ possible measurements. This implies in particular that adaptivity does not give us any advantage using single-copy measurements that are efficiently implementable. We also obtain a similar bound in the case where the goal is to predict the expectation values of a given sequence of observables, a task known as shadow tomography. Finally, in the case of adaptive, single-copy measurements implementable with polynomial-size circuits, we prove that a straightforward strategy based on computing sample means of the given observables is optimal.

1 Introduction

1.1 State tomography and its variants

In learning theory, an important resource is the number of samples of data used by the learner to correctly infer or predict their properties. The difficulty of a learning task, at first approximation, is therefore captured by its *sample complexity*, defined to be the minimum number of samples required to solve the problem at hand with high probability. In this paper we consider the sample complexity of learning properties of an arbitrary unknown quantum state. Here, a sample

*Most of this work was completed while this author was with Department of Combinatorics & Optimization and Institute for Quantum Computing, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada. Email: alowe7@mit.edu .

†Department of Combinatorics & Optimization and Institute for Quantum Computing, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada. Email: ashwin.nayak@uwaterloo.ca .

amounts to preparing the state in some register, so that the number of samples is the number of identical copies of the state on which the learner can perform a measurement. For the most part, we focus on quantum state tomography, which is the fundamental task of estimating an unknown d-dimensional state ρ to within some accuracy ϵ in the standard trace distance between states. Quantum tomography is of significant practical interest, for example, for the experimental verification of quantum devices. We are especially interested in how the sample complexity of tomography scales with the dimension d of the state. In theory, the dimension is the primary obstacle to efficient learning, since this quantity grows exponentially with the number of qubits comprising the system.

In the most general scenario for state tomography, n identical copies of a state ρ are prepared in registers that are jointly measured. It is then said that the measurements are *entangled*. In a series of breakthroughs, O'Donnell and Wright [OW16, OW17] as well as Haah, Harrow, Ji, Wu, and Yu [HHJ$^+$17] proved that $O(d^2/\epsilon^2)$ samples suffice to perform tomography using entangled measurements. This matches an information-theoretic lower bound due to Ref. [HHJ$^+$17] and improves upon previous upper bounds by a factor of d. Fewer samples are needed when a bound on the rank of the state is known (see, for instance, Ref. [OW16]).

From a practical standpoint, however, the joint measurements used in algorithms for optimal tomography may not be feasible. Firstly, in the case where one has access to just a single register that can be prepared in the state ρ, joint measurements of multiple copies of the state are impossible. (For instance, one might wish to perform tomography on the output state of a quantum computer by repeating a computation. Another example is that of photonic states that are difficult to store over extended periods of time.) Even given access to a suitably large system that can be prepared in the state $\rho^{\otimes n}$, it is not clear how efficiently the entangled measurements can be implemented. Finally, in some experimental realizations, only a limited set of measurements may be available. For these reasons, there is strong motivation to consider restricted measurement models, for instance, those in which each copy of ρ is measured separately, possibly using one of a fixed set of measurement settings. Measurements in which each copy of ρ is measured separately have been coined *single-copy* measurements by some [ACH$^+$19, ALL22] (and *unentangled* measurements by others [CD10, Wri16, BCL20]).

Within the single-copy model of measurement, one has access only to a single d-dimensional register which can be repeatedly prepared in the state ρ upon request, at which point a measurement is performed on the state and the resulting state is discarded. This means that the number of samples is equal to the number of measurements performed. Upper bounds on the sample complexity of single-copy tomography are well-established. Two prominent examples are the folklore "Pauli tomography algorithm" (outlined in Section 8.4.2 in Nielsen and Chuang [NC10]) and algorithms based on low-rank matrix recovery due to Kueng, Rauhut, and Terstiege [KRT17]. In both examples, the upper bound on the sample complexity is worse than in the entangled case. (For other, simple such algorithms, see Refs. [Wri16, GKKT20, Yu20].)

What can be said about the sample complexity of quantum tomography using single-copy measurements? Haah *et al.* [HHJ$^+$17] address this question by providing a $\Omega(d^3/\epsilon^2)$ lower bound which matches the upper bound following from Ref. [KRT17, GKKT20], under the assumption that the choice of each measurement is independent of any previous outcomes (referred to as *nonadaptive measurements*). However, this does not exhaust all realizable single-copy measurement strategies. Indeed, numerous proposals for state tomography (e.g., [HH12, MRD$^+$13]) utilize *adaptive* measurements, where the choice of measurement can depend on previous outcomes.

Adaptive measurements represent an intermediate restriction between nonadaptive and entangled measurements, and until very recently little was known about the sample complexity of learning quantum states or their properties in this setting. (For an early example of a prob-

lem for which adaptive measurements do not help, see Refs. [HRS05, HMR+10].) This is despite the fact that bounding the power of adaptivity is a significant problem: proving separations between entangled and single-copy measurements requires showing that adaptive measurements result in strictly worse sample complexity. It was posed as an open problem in the Ph.D. thesis of Wright [Wri16] to provide examples where this is the case, and since then there has been significant progress on this topic. In 2020, Bubeck, Chen, and Li [BCL20] gave the first unconditional separation between entangled and single-copy measurements, for the problem of quantum state certification. Following this, Huang, Kueng and Preskill [HKP21] proved an *exponential* separation for the problem of determining the expectations of Pauli operators to constant accuracy. Then, in 2021 Chen, Cotler, Huang, and Li [CCHL22, CCHL21] proved many additional exponential separations for different learning tasks, including shadow tomography. In this work, we continue along this line of research to investigate the sample complexity of adaptive quantum tomography in a realistic setting. We then apply the techniques developed and derive a new lower bound for single-copy shadow tomography in the same setting.

1.2 Summary of results

We first provide a simplified proof of the lower bound for tomography in the nonadaptive case due to Haah *et al.* [HHJ+17, Theorem 4]. In the process, we improve it by a factor of d to $\Omega(d^4/\epsilon^2)$ when the measurements have a constant number of outcomes. This implies that the straightforward Pauli tomography algorithm (described in Appendix B) is information-theoretically optimal in this setting. Using the same techniques, we derive a lower bound of $\Omega(r^2 d/\epsilon^2)$ when the states are known to have bounded rank r. This bound is a multiplicative factor of $\log(1/\epsilon)$ larger than the best previous lower bound [HHJ+17, Theorem 4], and is optimal [KRT17, GKKT20] (see Section B.2 for more details on the upper bound). Moreover, it applies to the case of learning pure states ($r = 1$), which is not covered by the proof of Theorem 4 in Ref. [HHJ+17]. The rank-dependent bound can be further strengthened to $\Omega(r^2 d^2/\epsilon^2)$ for measurements with a constant number of outcomes.

Since state tomography requires $\Omega(d^2/\epsilon^2)$ samples, any quantum algorithm for this problem necessarily has run-time at least quadratic in d. This is exponential in $\log d$, the number of qubits representing the unknown state. However, algorithms that measure one copy of the state at a time, interleaved with classical processing of the measurement outcomes, allow for the possibility that the *individual* measurements be more time-efficient. Such algorithms are more attractive from a practical point of view, given the current challenges in implementing quantum computation. It is thus no surprise that most of the algorithms based on single-copy measurements mentioned in Section 1.1 involve measurements that can be implemented efficiently, in particular with quantum circuits of size *polynomial* in the number of qubits.

We present new arguments showing there is a broad class of algorithms, including the ones described above, for which adaptivity *makes no difference* to the worst-case sample complexity of learning a quantum state. Specifically, we prove a lower bound of $\Omega(d^3/\epsilon^2)$ for the sample complexity of any single-copy, adaptive tomography algorithm which uses measurements chosen from a fixed set of up to $\exp(O(d))$ measurements. This encompasses measurement strategies which are efficiently implementable, i.e., the measurements may be performed using (uniformly generated) circuits of size polynomial in $\log d$ over some finite universal gate-set. We also show using the Solovay-Kitaev Theorem that, up to a factor of roughly $\log\log d + \log(1/\epsilon)$, the same bound applies to all measurement strategies which are efficiently implementable using circuits on possibly infinite universal gate-sets. The bounds entail that either (i) adaptivity does not give any advantage over non-adaptive measurements for single-copy tomography, or (ii) any adaptive algorithm using $o(d^3/\epsilon^2)$ samples necessarily uses measurements with super-polynomial-size cir-

	Nonadaptive		Adaptive	Adaptive & efficient	
Allowed meas.	$O(1)$-outcome	Arbitrary	Binary Pauli	$O(1)$-outcome	Arbitrary
Upper bound	$O(d^4/\epsilon^2)$	$O(d^3/\epsilon^2)$ [KRT17, GKKT20]	$O(d^4/\epsilon^2)$	$O(d^4/\epsilon^2)$	$O(d^3/\epsilon^2)$
Lower bound	$\Omega(d^4/\epsilon^2)$ [∗]	$\Omega(d^3/\epsilon^2)$ [HHJ+17]	$\Omega(d^4)$ [FGLE12]	$\widetilde{\Omega}(d^4/\epsilon^2)$ [∗]	$\widetilde{\Omega}(d^3/\epsilon^2)$ [∗]

Table 1: Best known upper and lower bounds for the sample complexity of quantum state tomography using single-copy measurements under various measurement restrictions, prior to this work. $\widetilde{\Omega}$ hides $\log(d)$ and $\mathrm{polylog}(1/\epsilon)$ factors, lack of citation indicates folklore or implied by other bounds, and [∗] denotes results from this work.

cuits. We summarize lower bounds for single-copy tomography in comparison to previous work in Table 1, in the full-rank case. In the final column, by "efficient" we mean efficiently implementable, as defined above.

We also obtain lower bounds of the above kind for computing classical shadows [HKP20] and for shadow tomography [ACH+19]. In these tasks, one is interested in estimating the expectations of some collection of observables, and they have practical applications ranging from entanglement verification to near-term proposals of variational quantum algorithms [HKP20, SZK+21]. We show that any procedure for ϵ-accurate shadow tomography of M observables using efficiently implementable single-copy measurements requires $\Omega(d\log(M)/\epsilon^2)$ samples of the unknown d-dimensional quantum state. Recently, Ref. [CCHL22] almost fully resolved the sample complexity of shadow tomography in the more general case where the learner can implement arbitrary single-copy measurements. They showed a lower bound of $\widetilde{\Omega}(\min\{M,d\}/\epsilon^2)$. This, while being more general than our result, is potentially exponentially looser in the setting of efficient measurements. In particular, even for M a small constant, our lower bound is linear in the dimension of the state, whereas the more general lower bound has no dependence on the dimension at all.

Finally, we present a simple procedure for shadow tomography using single-copy measurements that are efficiently implementable. The algorithm is optimal in this setting as well as in the case where the measurements are nonadaptive but otherwise arbitrary. The procedure is simpler than the one given in Ref. [HKP20].

Subsequent work. Most of the results in this article were included in the first author's Master's thesis [Low21] and were presented at QIP 2022 [LN22]. Chen, Huang, Li, and Liu [CHLL22] subsequently proved that known non-adaptive algorithms for state certification are optimal even when adaptive measurements are used. More recently, the same set of authors along with Sellke [CHL+22] reported an $\Omega(d^3/\epsilon^2)$ lower bound on the sample complexity of tomography of states of possibly full rank, using adaptive single-copy measurements. These bounds imply that adaptivity does not give any advantage over non-adaptive measurements in terms of sample complexity for state tomography (as a function of the dimension) or for the related tasks mentioned above.

1.3 Overview of techniques

We first describe a basic framework for proving lower bounds on the task of quantum tomography common to much of the work on the topic. Here, we use the observation that state discrimination of well-separated states reduces to tomography with sufficient accuracy. The lower bounds then follow from the construction of difficult instances of the state discrimination problem, for which the amount of information that the measurement statistics can reveal about the chosen state is

severely limited. "Discretizing" the learning problem in this manner for the purposes of providing worst-case lower bounds is a standard technique in the field of density estimation, which is the classical analogue of quantum tomography. (See for example Chapter 2 of Ref. [Tsy09].) To the best of the authors' knowledge, the method was first employed in the context of tomography by Flammia, Gross, Liu, and Eisert [FGLE12].

One way to make this argument rigorous is by using Fano's inequality and Holevo's theorem, which suggests an interpretation in terms of a communication protocol between two parties, Alice and Bob. To this end, imagine they have agreed upon an encoding of 2^N quantum states into bit-strings x of length N. In a single round of communication, Alice sends a quantum state $\rho_x^{\otimes n}$ encoding the message $x \in \{0,1\}^N$ to Bob who then attempts to decode the message through tomography. Assuming Bob can perform accurate tomography using n copies of the unknown state, Alice will have successfully transmitted N bits of information to Bob. On the other hand, the Holevo information of the ensemble of quantum states gives an upper bound on the size of a message that could be sent reliably. In particular, it can be shown that when n is small the Holevo information is also small. This provides the necessary contradiction to arrive at a lower bound: a procedure for tomography that succeeds when n is small could be used by Bob to reliably decode too long a message from Alice. Therefore, there is no such procedure.

In summary, this argument may be used to show that the mutual information between the random choice of state $\boldsymbol{x}$ and the measurement outcome $\boldsymbol{y}$ satisfies $\Omega(d^2) \leq I(\boldsymbol{x} : \boldsymbol{y}) \leq n\epsilon^2$, where the first inequality comes from Fano's inequality, and the second from Holevo's theorem. However, using Holevo's theorem in this manner does not take into account restrictions on the measurements we are allowed to perform on the n copies of the state. One might therefore expect that it be possible to derive a tighter bound on the mutual information by exploiting the fact that the measurements are not entangled. It turns out that this is indeed the case, as demonstrated by the $\Omega(d^3/\epsilon^2)$ lower bound for nonadaptive measurements due to Ref. [HHJ$^+$17].

Our approach differs from previous work in making direct use of a connection between the mutual information of two random variables and the χ^2-divergence of related distributions, as well as techniques for Haar integration based on symmetry. Additionally, we do not require that the measurements be rank-one POVMs as in Ref. [HHJ$^+$17]; this allows us to conclude the $\Omega(d^4/\epsilon^2)$ lower bound in the constant-outcome, nonadaptive case, as well as the more precise bounds stated in Section 1.2 for states of bounded rank. We further build on these simplifications to derive lower bounds robust to a wide class of adaptive measurements. We accomplish this by adversarially constructing instances of the state discrimination problem that are as difficult as possible for the specific set of measurements under consideration. This involves making use of well-known concentration of measure results for the unitary group. This idea is reminiscent of the lower bounds for tomography restricted to binary Pauli measurements due to Flammia *et al.* [FGLE12]. A key technical step is the analysis of χ^2-divergence rather than the probability of individual measurement outcomes. This enables tight lower bounds agnostic to the measurements we consider.

2 Preliminaries

2.1 Mathematical background

This section contains relevant notation and properties that may be referred to as needed.

Sets. We let $\mathbb{Z}_+$ denote the set of nonnegative integers, $\mathsf{U}(d)$ the set of unitary operators acting on $\mathbb{C}^d$, $\mathsf{H}(d)$ the set of Hermitian operators acting on $\mathbb{C}^d$, $\mathsf{Psd}(d)$ the subset of $\mathsf{H}(d)$ consisting of positive semidefinite operators, and $\mathsf{D}(d)$ the subset of operators in $\mathsf{Psd}(d)$ with unit trace (i.e., the set of d-dimensional quantum states). We also denote by $\mathsf{L}(d)$ the set of square operators acting on $\mathbb{C}^d$.

Operators. For any square operator $A \in \mathsf{L}(d)$, we denote its adjoint by $A^\dagger$. We let $\|A\|_1 = \mathrm{Tr}(\sqrt{A^\dagger A})$ denote the "trace norm" of the operator A and $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^\dagger A)}$ it Frobenius norm. The *trace distance* between two quantum states is $\|\rho - \sigma\|_1$. We use $\|A\|$ to denote the spectral norm of the operator A; this is the operator norm induced by the Euclidean norm on $\mathbb{C}^d$. We have the useful relations $\|A\|_{\mathrm{F}} \leq \|A\|_1 \leq \sqrt{d}\,\|A\|_{\mathrm{F}}$ and $\|AB\|_{\mathrm{F}} \leq \|A\|\,\|B\|_{\mathrm{F}}$. For any two operators $P, Q \in \mathsf{Psd}(d)$, we use the notation $P \preceq Q$ if and only if $Q - P \in \mathsf{Psd}(d)$. Let $A, B \in \mathsf{H}(d)$ and consider the operator $A \otimes B$. We denote by $\mathrm{Tr}_2(\cdot)$ the partial trace over the second system, i.e., $\mathrm{Tr}_2(A \otimes B) = A\,\mathrm{Tr}(B)$. The rank of a linear operator X, denoted $\mathrm{rank}(X)$, is the dimension of its image, which we denote by $\mathrm{im}(X)$.

Permutation operator and t-designs. The swap operator W acting on $(\mathbb{C}^d)^{\otimes 2}$ is the linear operator defined by the action $W|\psi\rangle \otimes |\phi\rangle = |\phi\rangle \otimes |\psi\rangle$ for any two vectors $|\psi\rangle, |\phi\rangle \in \mathbb{C}^d$. We may extend this procedure to arbitrary permutations, defining the linear operator W_π for each $\pi \in S_n$ and acting on $(\mathbb{C}^d)^{\otimes n}$ as

$$W_\pi \, |x_1\rangle \otimes \cdots \otimes |x_n\rangle = |x_{\pi^{-1}(1)}\rangle \otimes \cdots \otimes |x_{\pi^{-1}(n)}\rangle$$

for every choice of vectors $|x_1\rangle, \ldots, |x_n\rangle \in \mathbb{C}^d$. Here, S_n denotes the symmetric group on $\{1, \ldots, n\}$.

We make use of unitary and state t-designs throughout this paper.

Definition 2.1 (Unitary t-design)**.** For positive integers $t, d > 0$ we say that a random unitary operator $\boldsymbol{U} \in \mathsf{U}(d)$ is a *unitary t-design* if the the following holds for every operator $X \in \mathsf{L}(d)^{\otimes t}$:

$$\mathbb{E}\, \boldsymbol{V}^{\otimes t} X (\boldsymbol{V}^\dagger)^{\otimes t} = \int_{\mathsf{U}(d)} U^{\otimes t} X (U^\dagger)^{\otimes t} \mathrm{d}\mu(U)$$

where μ is the Haar measure on the space of d-dimensional unitary operators.

Definition 2.2 (State t-design)**.** For positive integers $t, d > 0$, a *state t-design* is a random quantum state $|\boldsymbol{u}\rangle \in \mathsf{S}(d)$ which satisfies

$$\mathbb{E}(|\boldsymbol{u}\rangle\langle\boldsymbol{u}|)^{\otimes t} = \int_{\mathsf{S}(d)} (|v\rangle\langle v|)^{\otimes t} \, \mathrm{d}\mu(v) \tag{1}$$

where $\mathsf{S}(d)$ is the set of unit vectors in $\mathbb{C}^d$.

Random variables. We denote random variables using bold font, including matrix-valued random variables. We use lowercase (e.g., p, q) with appropriate subscripts to denote the distributions of random variables. For example, suppose $\boldsymbol{x}$ is a random variable taking values in $\mathcal{X}$ according to some distribution $p_{\boldsymbol{x}} : \mathcal{A} \to [0,1]$, where $\mathcal{A}$ is the set of Borel-measurable subsets of $\mathcal{X}$. Let $\mathcal{S}$ be some finite-dimensional vector space, and let $f : \mathcal{X} \to \mathcal{S}$. Then we write interchangeably $\mathbb{E}_{\boldsymbol{x}} f(\boldsymbol{x})$ and $\mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} f(\boldsymbol{x})$ to refer to the expectation of f with respect to the distribution $p_{\boldsymbol{x}}$ (i.e., $\int_{\mathcal{X}} f(x) \mathrm{d}p_{\boldsymbol{x}}(x)$) using the latter notation when there may be some ambiguity about what the distribution is. When it is clear enough from context, we drop the subscripts altogether and write

$\mathbb{E} f(\boldsymbol{x})$. In the case where $\boldsymbol{x}$ is a discrete random variable taking values in some finite set (or alphabet) $\mathcal{X}$, we write its probability mass function (PMF) as $p_{\boldsymbol{x}}$, and corresponding expectations $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} f(\boldsymbol{x}) = \sum_{x\in\mathcal{X}} p_{\boldsymbol{x}}(x) f(x)$. We also refer to $p_{\boldsymbol{x}}$ as the distribution of $\boldsymbol{x}$ in this case. Next suppose we have random variables $(\boldsymbol{x}, \boldsymbol{y})$ jointly distributed on $\mathcal{X} \times \mathcal{Y}$. If $\boldsymbol{y}$ is discrete, we write $p_{\boldsymbol{y}|x}(y)$ to mean the probability that $\boldsymbol{y} = y$ given $\boldsymbol{x} = x$, when it is well-defined. We will often have occasion to use functionals F mapping distributions to the reals. Then if $\boldsymbol{x}$ has marginal distribution given by $p_{\boldsymbol{x}}$, we write $\mathbb{E}_{\boldsymbol{x}'\sim p_{\boldsymbol{x}}} F(p_{\boldsymbol{y}|\boldsymbol{x}'})$ to denote the expectation $\int_{\mathcal{X}} F(p_{\boldsymbol{y}|x}) \mathrm{d}p_{\boldsymbol{x}}(x)$. Finally, we sometimes use in the subscripts of expectations the notation $\boldsymbol{x}|y$ to mean the random variable $\boldsymbol{x}$ conditioned on $\boldsymbol{y} = y$, when it is well-defined. For example, suppose we have a function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. It holds by definition that $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{y}} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y})$.

Information theory. Consider discrete random variables taking values on the same space. One may then use the KL-divergence between their distributions to compare them. The KL-divergence between two discrete distributions (PMFs) $p, q : \mathcal{X} \to [0,1]$ defined on the same sample space $\mathcal{X}$ is

$$\mathrm{D}_{\mathrm{KL}}(p \parallel q) = \begin{cases} \sum_{x\in\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right), & \mathrm{supp}(p) \subseteq \mathrm{supp}(q) \\ +\infty, & \text{otherwise} \end{cases}$$

where we take $0\log(0) = 0$. (Throughout this work, log denotes the logarithm with base 2.)

We next define some entropic quantities. Let $\boldsymbol{x}$ be a discrete random variable taking values in $\mathcal{X}$ with distribution $p_{\boldsymbol{x}} : \mathcal{X} \to [0,1]$. The Shannon entropy measures our uncertainty about $\boldsymbol{x}$ and is defined as

$$H(\boldsymbol{x}) = -\sum_{x\in\mathcal{X}} p_{\boldsymbol{x}}(x) \log(p_{\boldsymbol{x}}(x)).$$

We also write $H(p_{\boldsymbol{x}})$ to refer to the same quantity. A useful property of the entropy is *concavity*, whereby for any two discrete distributions p, q defined on the same sample space and $\lambda \in [0,1]$ it holds that

$$H(\lambda p + (1-\lambda)q) \geq \lambda H(p) + (1-\lambda) H(q).$$

Next, let $\boldsymbol{y}$ be a different discrete random variable taking values in $\mathcal{Y}$, so that $\boldsymbol{x}$ and $\boldsymbol{y}$ have joint distribution given by $p_{\boldsymbol{x},\boldsymbol{y}} : \mathcal{X} \times \mathcal{Y} \to [0,1]$. The joint entropy of these random variables is

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{\boldsymbol{x},\boldsymbol{y}}(x,y) \log(p_{\boldsymbol{x},\boldsymbol{y}}(x,y))$$

and the conditional entropy of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$H(\boldsymbol{x}|\boldsymbol{y}) = H(\boldsymbol{x}, \boldsymbol{y}) - H(\boldsymbol{y}).$$

These definitions are valid only in the case where $\boldsymbol{x}$ and $\boldsymbol{y}$ are discrete. Mutual information, on the other hand, is well-defined for arbitrary random variables $\boldsymbol{x}, \boldsymbol{y}$ though for our purposes it will suffice to define this quantity in the following way, which is valid when $\boldsymbol{y}$ is discrete.

Definition 2.3 (Mutual information)**.** Consider two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $\boldsymbol{y}$ is discrete. Let $p_{\boldsymbol{y}|x}$ be the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x} = x$, $p_{\boldsymbol{x}}$ the marginal distribution of $\boldsymbol{x}$, and $p_{\boldsymbol{y}}$ the marginal distribution of $\boldsymbol{y}$. The *mutual information* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$I(\boldsymbol{x} : \boldsymbol{y}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel p_{\boldsymbol{y}}).$$

As the name suggests, the mutual information between two random variables quantifies the shared information between them. Since this definition is somewhat non-standard, it is worth taking the time to see how it reduces to the more standard definitions in familiar settings. Firstly, it may be shown that the above is equal to

$$I(\boldsymbol{x}:\boldsymbol{y}) = H(\boldsymbol{y}) - \mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_{\boldsymbol{x}}} H(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}')$$

where $\boldsymbol{y}|\boldsymbol{x}=x$ is the random variable $\boldsymbol{y}$ conditioned on the event $\boldsymbol{x}=x$. Then, if $\boldsymbol{x}$ is also discrete, it holds that $H(\boldsymbol{y}|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x}'\sim p_{\boldsymbol{x}}} H(\boldsymbol{y}|\boldsymbol{x}=\boldsymbol{x}')$ in which case we arrive at the commonly used expression for the mutual information,

$$I(\boldsymbol{x}:\boldsymbol{y}) = H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) = H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y}) = H(\boldsymbol{x}) + H(\boldsymbol{y}) - H(\boldsymbol{x},\boldsymbol{y}).$$

Next, suppose $\boldsymbol{z}$ is another random variable jointly distributed with $\boldsymbol{x}$ and $\boldsymbol{y}$. When $\boldsymbol{z}$ has a fixed value z, we use the notation

$$I(\boldsymbol{x}:\boldsymbol{y}|\boldsymbol{z}=z) \coloneqq I(\ (\boldsymbol{x}|\boldsymbol{z}=z)\ :\ (\boldsymbol{y}|\boldsymbol{z}=z)\ )$$

where $(\boldsymbol{x}|\boldsymbol{z}=z)$ is $\boldsymbol{x}$ conditioned on $\boldsymbol{z}=z$, and likewise for $(\boldsymbol{y}|\boldsymbol{z}=z)$. The conditional mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ given $\boldsymbol{z}$ is then defined as

$$I(\boldsymbol{x}:\boldsymbol{y}|\boldsymbol{z}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{z}'\sim p_{\boldsymbol{z}}} I(\boldsymbol{x}:\boldsymbol{y}|\boldsymbol{z}=\boldsymbol{z}').$$

We now present three exceedingly useful facts about mutual information. We will use these to derive stronger lower bounds on tomography than the ones obtained by applying Holevo's theorem in the case where there is some restriction on the measurements.

Fact 2.4 (Chain rule for mutual information)**.** *It holds that*

$$I(\boldsymbol{x}:\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) = \sum_{i=1}^{n} I(\boldsymbol{x}:\boldsymbol{y}_i|\boldsymbol{y}_{i-1},\ldots,\boldsymbol{y}_1).$$

Corollary 2.5 (Subadditivity of mutual information)**.** *If $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are independent given $\boldsymbol{x}$, it holds that*

$$I(\boldsymbol{x}:\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) \leq \sum_{i=1}^{n} I(\boldsymbol{x}:\boldsymbol{y}_i).$$

The random variables $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ form a *Markov chain* $\boldsymbol{x}\to\boldsymbol{y}\to\boldsymbol{z}$ if given $\boldsymbol{y}$, the random variables $\boldsymbol{x}$ and $\boldsymbol{z}$ are independent (Ref. [CT05], Section 2.8). Under this assumption, the following lemma holds, which is indispensable toward proving information-theoretic lower bounds on estimation tasks.

Lemma 2.6 (Fano's inequality [Fan66])**.** *Let $\boldsymbol{x}$, $\boldsymbol{y}$, $\hat{\boldsymbol{x}}$ be discrete random variables forming a Markov chain $\boldsymbol{x}\to\boldsymbol{y}\to\hat{\boldsymbol{x}}$, where $\boldsymbol{x}$ takes values in $\mathcal{X}$. It holds that*

$$H(p_e) + p_e \log(|\mathcal{X}|) \geq H(\boldsymbol{x}|\boldsymbol{y}).$$

where $p_e \coloneqq \Pr[\boldsymbol{x}\neq\hat{\boldsymbol{x}}]$, and $H(\cdot)$ is the binary entropy function.

Corollary 2.7. *Let $\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{x}}$ be discrete random variables forming a Markov chain $\boldsymbol{x} \to \boldsymbol{y} \to \hat{\boldsymbol{x}}$. Suppose Alice has a message $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and Bob is able to decode the message with constant probability of success using $\hat{\boldsymbol{x}}$. It must hold that*

$$I(\boldsymbol{x} : \boldsymbol{y}) = \Omega(\log(N)).$$

Proof. Using the definition of mutual information we have $I(\boldsymbol{x} : \boldsymbol{y}) = H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y})$. Let p_e be as in Lemma 2.6. By Lemma 2.6 we have $I(\boldsymbol{x} : \boldsymbol{y}) \geq H(\boldsymbol{x}) - p_e \log(N) - H(p_e)$. Using the fact that $H(\boldsymbol{x}) = \log(N)$ and $H(p_e) \leq 1$ we obtain $I(\boldsymbol{x} : \boldsymbol{y}) \geq (1 - p_e)\log(N) - 1$. □

Besides the KL-divergence, another way to compare distributions defined on the same space is the following.

Definition 2.8 (χ^2-divergence)**.** The χ^2-divergence between two discrete distributions $p, q : \mathcal{X} \to [0,1]$ defined on the same sample space $\mathcal{X}$ is

$$\mathrm{D}_{\chi^2}(p \parallel q) := \sum_{x \in \mathcal{X}} q(x) \left(\frac{p(x)}{q(x)} - 1 \right)^2 = \sum_{x \in \mathcal{X}} q(x) \left(\frac{p(x)}{q(x)} \right)^2 - 1.$$

These divergences are related in the following way.

Lemma 2.9 (KL vs. χ^2 inequality)**.** *Let $p, q : \mathcal{X} \to [0,1]$ be discrete distributions defined on the same sample space $\mathcal{X}$. We have*

$$\mathrm{D}_{\mathrm{KL}}(p \parallel q) \leq \frac{1}{\ln(2)} \cdot \mathrm{D}_{\chi^2}(p \parallel q).$$

Proof. By Eq. (5) in Ref. [SV16], we have the inequality $\mathrm{D}_{\mathrm{KL}}(p \parallel q) \leq \log(1 + \mathrm{D}_{\chi^2}(p \parallel q))$ from which the lemma follows by the inequality $\log(1 + x) \leq x / \ln(2)\ \forall x \geq 0$. For an exposition of the many other relationships between divergences, we refer the interested reader to Ref. [SV16]. □

2.2 Single-copy measurements

In general, an m-outcome *measurement* of a d-dimensional quantum state is a linear map $\mathcal{M} : \mathsf{D}(d) \to \mathsf{L}(m)$ acting on quantum states $\rho \in \mathsf{D}(d)$ by

$$\mathcal{M} : \rho \mapsto \sum_{z \in \mathcal{Z}} \mathrm{Tr}(M_z \rho) |z\rangle\langle z|$$

for some "positive operator-valued measure" (POVM) $(M_z : z \in \mathcal{Z},\ M_z \in \mathsf{Psd}(d))$ satisfying $\sum_{z \in \mathcal{Z}} M_z = \mathbb{1}$, and where $\mathcal{Z}$ is a set of m possible outcomes of the measurement. For a measurement $\mathcal{M}$, $\mathrm{rank}(\mathcal{M})$ denotes the number of possible outcomes $|\mathcal{Z}|$. Without loss of generality we can assume $\mathcal{Z} = [m]$. In this work we focus on measurements with a finite number of outcomes, letting $\Xi(d, m)$ denote the set of all m-outcome measurements on d-dimensional states, and $\Xi(d) := \bigcup_{m \in \mathbb{Z}_+} \Xi(d, m)$ denote the set of all finite-outcome measurements on d-dimensional states. The distribution of the random outcome $\boldsymbol{z}$ from measuring the state ρ is described by the PMF $p_{\boldsymbol{z}} = \mathrm{diag}(\mathcal{M}(\rho))$, so that $p_{\boldsymbol{z}}(z) = \mathrm{Tr}(M_z \rho)$ for all outcomes z.

Suppose there is a single d-dimensional register which can be prepared in the state ρ upon request, at which point it is measured once, and this process is repeated n times. We refer to the class of measurements corresponding to this scenario as *single-copy measurements*, where the number of samples used is equal to the number of measurements performed. Within this class, there are two models of particular interest.

Nonadaptive measurements. Consider n copies of the state $\rho \in \mathsf{D}(d)$ prepared in the above manner, so that they must be measured individually. In the *nonadaptive* measurement model, we use a sequence of measurements $\mathcal{M}_i \in \Xi(d)$ for $i = 1, \ldots, n$ which are determined before any measurements are performed. Equivalently, we measure the state $\rho^{\otimes n}$ using a tensor product of measurements on d-dimensional states, $\mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$. Note that allowing the choice of the i^{th} measurement to be an independent random variable is equivalent to the above description, since the randomness in the choice of measurement can be incorporated into the measurement itself. I.e., the resulting linear maps on d-dimensional states still correspond to some fixed measurements.

Adaptive measurements. In the *adaptive* measurement model, the choice of each d-dimensional measurement in the sequence can depend on the outcomes obtained by the previous measurements. This means that the i^{th} measurement in the sequence can be written $\mathcal{M}^{y_{<i}}$, where $y_{<i} = y_{i-1} \ldots y_1$ are the outcomes of the previous $i-1$ measurements. For each possible value of $y_{<i}$ there is a POVM $(M_{y_i}^{y_{<i}})_{y_i}$ corresponding to i^{th} measurement and potentially depending on $y_{<i}$, such that the measurement has the action

$$\mathcal{M}^{y_{<i}} : \rho \mapsto \sum_{y_i} \mathrm{Tr}(M_{y_i}^{y_{<i}} \rho) |y_i\rangle\langle y_i|$$

on quantum states $\rho \in \mathsf{D}(d)$.

3 Packing construction

To demonstrate lower bounds for quantum tomography, it suffices to show that there exists a large, but well-separated collection of quantum states (an ϵ-packing) which are difficult to discriminate with too few copies of the state. This is due to the fact that the task of state discrimination reduces to tomography with sufficient accuracy when the states are far enough apart, since the latter task allows one to correctly identify the state in the ensemble under these conditions. We therefore aim to construct a hard instance of the state discrimination problem, and then argue that if the number of samples n is too small the success probability of our protocol goes to zero as the parameters d and $1/\epsilon$ increase.

Definition 3.1 (ϵ-packing)**.** A finite set of quantum states $\mathcal{S} \subset \mathsf{D}(d)$ is an *ϵ-packing* for some $\epsilon > 0$ if it holds that $\|\rho - \sigma\|_1 > \epsilon$ for every $\rho, \sigma \in \mathcal{S}$ such that $\rho \neq \sigma$.

Let $\{|i\rangle : i \in [d]\}$ denote the standard basis for $\mathbb{C}^d$, and let Q_k be the orthogonal projection operator onto the subspace spanned by $\{|i\rangle : i \in [k]\}$. The ϵ-packing we construct comprises states of the following form:

$$\rho_{\epsilon,U} := \frac{2\epsilon}{d} U Q_{d/2} U^\dagger + \frac{1-\epsilon}{d} \mathbb{1} \tag{2}$$

where $\epsilon \in (0,1)$ and we assume d is even for simplicity. The assumption of d being even does not take away from the argument, and we may proceed analogously with a floor or ceiling when it is odd. States of the above form have also been considered in the previous lower bounds for tomography and related tasks (see, e.g., Refs. [HHJ+17, BCL20]). Intuitively, these states are useful because they represent a hard case where the completely mixed state is slightly perturbed, which leads to "noisy" measurement statistics. This is in analogy with the packing of distributions which one would construct to prove lower bounds for distribution estimation, the classical analogue of tomography. We make use of the definition in Eq. (2) frequently in the remainder of this paper.

We apply the probabilistic method to construct an ϵ-packing of states of this form. We draw a sequence of i.i.d. unitary operators $U_1, U_2, \ldots$ from the Haar distribution on $\mathsf{U}(d)$ and consider the states ρ_{ϵ,U_i}. We then apply standard concentration of measure results to argue that the probability of selecting an undesirable state (that our state "collides" with a previously chosen one) is exponentially small. This in turn implies that a large fraction of the states are "safe" choices, so that we may choose one and repeat the argument many times.

We use the following "concentration of projector overlaps" result, which is implied by the proof of Lemma III.5 in Ref. [HLW06], and has also been employed in the lower bounds for tomography which appear in [HHJ+17] as well as lower bounds for similar tasks (see for example Refs. [Aar20, HKP21]).

Lemma 3.2. *Let $\boldsymbol{U} \in \mathsf{U}(d)$ be a Haar-random unitary operator and let $\Pi_1, \Pi_2 \in \mathsf{Psd}(d)$ be orthogonal projection operators with rank r_1,r_2 respectively. For all $t \in (0,1)$ it holds that*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{Tr}(\Pi_1 \boldsymbol{U} \Pi_2 \boldsymbol{U}^\dagger) \leq (1-t)\frac{r_1 r_2}{d}\right] \leq \exp\left(-r_1 r_2 t^2/2\right)$$

$$\text{and} \quad \Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{Tr}(\Pi_1 \boldsymbol{U} \Pi_2 \boldsymbol{U}^\dagger) \geq (1+t)\frac{r_1 r_2}{d}\right] \leq \exp\left(-r_1 r_2 t^2/4\right) \quad .$$

Proof. By Lemma III.5 in Ref. [HLW06] we have

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{Tr}(\Pi_1 \boldsymbol{U} \Pi_2 \boldsymbol{U}^\dagger) \leq (1-t)\frac{r_1 r_2}{d}\right] \leq \exp\left(r_1 r_2 (t + \ln(1-t))\right)$$

for all $t \in (0,1)$, and the first bound follows immediately from the inequalities $\ln(1-t) \leq -t - t^2/2$ which holds for all $t \in (0,1)$. Similarly, the second bound in Lemma III.5 of Ref. [HLW06] is

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{Tr}(\Pi_1 \boldsymbol{U} \Pi_2 \boldsymbol{U}^\dagger) \geq (1+t)\frac{r_1 r_2}{d}\right] \leq \exp\left(-r_1 r_2 (t - \ln(1-t))\right) \tag{3}$$

for all $t \in (0,1)$, and noting that the inequality $\ln(1+t) \leq t - t^2/4$ holds for all $t \in (0,1)$ completes the proof. □

The second tail bound above is a bit looser than that shown in Ref. [HHJ+17], but suffices for our purposes. We now construct a sufficiently large packing of quantum states of the form in Eq. (2) which are difficult to discriminate, using a probabilistic existence argument. This is a special case of the approach adopted in Ref. [HHJ+17].

Lemma 3.3. *Fix an $\epsilon \in (0,1)$ and a positive integer d, and let $N \leq \lfloor \xi \mathrm{e}^{d^2/32} \rfloor$ be a positive integer for some $\xi \in (0,1]$. Consider a finite set of quantum states $\{\rho_1, \rho_2, \ldots, \rho_N\} \subset \mathsf{D}(d)$ where*

$$\rho_i = \frac{2\epsilon}{d} U_i Q_{d/2} U_i^\dagger + (1-\epsilon)\frac{\mathbb{1}}{d}$$

for each $i \in [N]$ and $U_1, U_2, \ldots, U_N \in \mathsf{U}(d)$ are arbitrary unitary operators. For Haar-random $\boldsymbol{U} \in \mathsf{U}(d)$, the probability that $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$ for any $i \in [N]$ is at most ξ.

Proof. Define a rank-$d/2$ orthogonal projection operator $P \in \mathsf{Psd}(d)$ as $P = \mathbb{1} - Q_{d/2}$. A straightforward consequence of Lemma 3.2 is that

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{Tr}(P \boldsymbol{U} Q_{d/2} \boldsymbol{U}^\dagger) \leq d/8\right] \leq \mathrm{e}^{-d^2/32}. \tag{4}$$

This follows by taking $t = 1/2$ in the lemma. Using the definition of $\rho_{\epsilon,U}$ we have

$$\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{1}} = \frac{2\epsilon}{d}\left(UQ_{d/2}U^\dagger - Q_{d/2}\right)$$

for any $U \in \mathsf{U}(d)$. We also have

$$\begin{aligned}
\operatorname{Tr}(PUQU^\dagger) &= \frac{1}{2}\left[\operatorname{Tr}(PUQ_{d/2}U^\dagger) + \operatorname{Tr}((\mathbb{1} - Q_{d/2})UQ_{d/2}U^\dagger)\right] && \text{(by the definition of } P\text{)} \\
&= \frac{1}{2}\operatorname{Tr}\left((UQ_{d/2}U^\dagger - Q_{d/2})(P - Q_{d/2})\right) && (P, Q_{d/2} \text{ are orthogonal}) \\
&\leq \frac{1}{2}\left\|UQ_{d/2}U^\dagger - Q_{d/2}\right\|_1 = \frac{d}{4\epsilon}\left\|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{1}}\right\|_1
\end{aligned}$$

where the final line follows from the property that $\|X\|_1 = \max\{|\operatorname{Tr}(XU)| : U \in \mathsf{U}(d)\}$ for any square operator $X \in \mathsf{L}(d)$, and $P - Q_{d/2} \in \mathsf{U}(d)$. Therefore, if $\|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{1}}\|_1 \leq \epsilon/2$ for a unitary operator $U \in \mathsf{U}(d)$, we also have that $\operatorname{Tr}(PUQ_{d/2}U^\dagger) \leq d/8$, from which we may conclude

$$\Pr_{\boldsymbol{U}\sim\text{Haar}}\left[\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{1}}\right\|_1 \leq \epsilon/2\right] \leq \mathrm{e}^{-d^2/32}$$

by Eq. (4). Next, consider the unitary operator U_i and corresponding state ρ_i in the lemma, for some $i \in [N]$. Using the invariance of the trace distance under unitary transformations, we have

$$\left\|\rho_{\epsilon,U_i\boldsymbol{U}} - \rho_i\right\|_1 = \left\|U_iUQ_{d/2}U^\dagger U_i^\dagger - U_iQ_{d/2}U_i^\dagger\right\|_1 = \left\|U_i(UQ_{d/2}U^\dagger - Q_{d/2})U_i^\dagger\right\|_1 = \left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{1}}\right\|_1$$

which leads to the conclusion that

$$\Pr_{\boldsymbol{U}\sim\text{Haar}}\left[\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\right\|_1 \leq \epsilon/2\right] \leq \mathrm{e}^{-d^2/32} \tag{5}$$

by invariance of the Haar measure. Since this inequality holds for any index $i \in [N]$ the proof is complete upon applying the union bound over the events $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$, $i \in [N]$: we have that this probability is at most $N\mathrm{e}^{-d^2/32} \leq \xi$. □

Using Lemma 3.3 we may construct a (non-explicit) set of N states with $N \in \exp(\Omega(d^2))$, which form an $\epsilon/2$-packing in trace distance, using a probabilistic existence argument.

Corollary 3.4. *Fix an $\epsilon \in (0,1)$ and a positive integer $d > 1$. There exists an $\epsilon/2$-packing $\mathcal{S} \subset \mathsf{D}(d)$ of $N \in \exp\left(\Omega(d^2)\right)$ quantum states of the form in Eq. (2).*

Proof. First, suppose we have a set of states $\mathcal{S}_k = \{\rho_1, \ldots, \rho_k\} \subset \mathsf{D}(d)$ which are of the same form as in Eq. (2), where $k \leq \lceil \mathrm{e}^{d^2/32} \rceil - 1$. Suppose further that this set is an $\epsilon/2$-packing. From Lemma 3.3 we know that the probability of choosing a unitary operator $\boldsymbol{U} \in \mathsf{U}(d)$ Haar randomly such that $\mathcal{S}_k \cup \{\rho_{\epsilon,\boldsymbol{U}}\}$ is *not* an $\epsilon/2$-packing is strictly less than one. Therefore, there exists at least one state which we can add to the packing. The result follows by induction on k. □

This packing of states is used in the following section to prove lower bounds for nonadaptive tomography. Then, in Section 5 we alter this construction to derive lower bounds on adaptive tomography.

4 Lower bounds for tomography with nonadaptive measurements

4.1 Information in measurement outcomes

We begin with some useful results quantifying our intuition that measurements performed on states in the packing described above are uninformative. Recall that in the nonadaptive case, measurement choices do not depend on the previously observed outcomes. The following lemma enables us to bound mutual information in terms of the χ^2-divergence, which is more amenable to analysis in this context.

Lemma 4.1. *Let $\boldsymbol{x}$ be an arbitrary random variable and $\boldsymbol{y} \in \mathcal{Y}$ be a discrete random variable for some sample space $\mathcal{Y}$. Denote by $p_{\boldsymbol{y}|x} : \mathcal{Y} \to [0,1]$ the distribution of $\boldsymbol{y}$ conditioned on the event $\boldsymbol{x} = x$. For an arbitrary discrete distribution $q : \mathcal{Y} \to [0,1]$, it holds that*

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq \frac{1}{\ln(2)} \mathop{\mathbb{E}}_{x \sim p_x} \mathrm{D}_{\chi^2}(p_{\boldsymbol{y}|x} \parallel q). \tag{6}$$

Proof. By Lemma 2.9 we have the inequality $\mathrm{D}_{\mathrm{KL}}(a \parallel b) \leq \mathrm{D}_{\chi^2}(a \parallel b)/\ln(2)$ for any two discrete distributions a and b defined on the same sample space. This implies the relation in Eq. (6) upon showing that

$$I(\boldsymbol{x} : \boldsymbol{y}) = \mathop{\mathbb{E}}_{x \sim p_x} \mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}|x} \parallel p_{\boldsymbol{y}}) \leq \mathop{\mathbb{E}}_{x \sim p_x} \mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}|x} \parallel q). \tag{7}$$

This inequality is a special case of Lemma 6 in Ref. [BD10], but for completeness we include a proof below. Using the definition of KL-divergence, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim p_x} \mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}|x} \parallel q) &= \mathop{\mathbb{E}}_{x \sim p_x} \sum_{y \in \mathcal{Y}} p_{\boldsymbol{y}|x}(y) \log\left(\frac{p_{\boldsymbol{y}|x}(y)}{q(y)}\right) \\
&= \sum_{y \in \mathcal{Y}} p_{\boldsymbol{y}}(y) \log\left(\frac{1}{q(y)}\right) - H(\boldsymbol{y}|\boldsymbol{x}) \\
&= \mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}} \parallel q) + H(\boldsymbol{y}) - H(\boldsymbol{y}|\boldsymbol{x}) \\
&= \mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}} \parallel q) + I(\boldsymbol{x} : \boldsymbol{y})
\end{aligned}$$

which proves the inequality in Eq. (7) since $\mathrm{D}_{\mathrm{KL}}(p_{\boldsymbol{y}} \parallel q) \geq 0$. □

Corollary 4.2. *Define $\boldsymbol{x}, \boldsymbol{y}$ as in Lemma 4.1. It holds that*

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq \frac{1}{\ln(2)} \mathop{\mathbb{E}}_{x \sim p_x} \mathrm{D}_{\chi^2}(p_{\boldsymbol{y}|x} \parallel p_{\boldsymbol{y}}) = \frac{1}{\ln(2)} \left(\sum_{y \in \mathcal{Y}} \mathop{\mathbb{E}}_{x \sim p_x} \frac{p_{\boldsymbol{y}|x}(y)^2}{p_{\boldsymbol{y}}(y)} - 1 \right). \tag{8}$$

In the analysis of state tomography, $\boldsymbol{x}$ corresponds to a random state from a suitably chosen ensemble. Although these results could be applied directly to the information contained in each measurement about $\boldsymbol{x}$, it would be intractable to compute an expectation over $\boldsymbol{x}$ since we do not explicitly know the states in our ensemble, whose existence is argued by means of the probabilistic method. Fortunately, we can make use of an intermediate result to effectively replace that ensemble with one which admits such explicit calculations, as explained in the following proposition. (A similar property is also used in the proof of Lemma 10 in Ref. [HHJ+17].)

Proposition 4.3. *Fix an $\epsilon \in (0,1)$ and a positive integer $d > 1$. Let $\boldsymbol{U} \in \mathsf{U}(d)$ be a Haar-random unitary operator and $\boldsymbol{z}$ be the outcome obtained upon measuring the random state $\rho_{\epsilon,\boldsymbol{U}}^{\otimes n} \in \mathsf{D}(d^n)$ with the measurement $\mathcal{M} \in \Xi(d^n)$, where $\rho_{\epsilon,U} \in \mathsf{D}(d)$ is defined as in Eq. (2) for any $U \in \mathsf{U}(d)$. There exists a set of $N \in \exp(\Omega(d^2))$ quantum states $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ of the form in Lemma 3.3 which is an $\epsilon/2$-packing and which satisfies*

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq I(\boldsymbol{U} : \boldsymbol{z})$$

where $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and $\boldsymbol{y}$ is the outcome obtained from measuring the random state $\rho_{\boldsymbol{x}}^{\otimes n}$ with $\mathcal{M}$.

Proof. Consider a fixed set of $N \in \exp(\Omega(d^2))$ quantum states $\mathcal{S}' = \{\rho'_1, \ldots, \rho'_N\} \subset \mathsf{D}(d)$ of the form in Lemma 3.3 which is an $\epsilon/2$-packing. We know such a set exists from Corollary 3.4. Let $\mathcal{U} = \{U_1, \ldots, U_N\}$ be the set of unitary operators such that $\rho'_i = \rho_{\epsilon,U_i}$ for each $i \in [N]$. Note that making the replacement $\mathcal{U} \to W\mathcal{U}$ for an arbitrary unitary operator $W \in \mathsf{U}(d)$ results in another $\epsilon/2$-packing of N states. Indeed, for any $\rho'_i, \rho'_j \in \mathcal{S}'$ we have

$$\begin{aligned}
\left\| \rho'_i - \rho'_j \right\|_1 &= \frac{2\epsilon}{d} \| U_i Q_{d/2} U_i^\dagger - U_j Q_{d/2} U_j^\dagger \|_1 \\
&= \frac{2\epsilon}{d} \| W U_i Q_{d/2} U_i^\dagger W^\dagger - W U_j Q_{d/2} U_j^\dagger W^\dagger \|_1 \\
&= \left\| \rho_{\epsilon,WU_i} - \rho_{\epsilon,WU_j} \right\|_1
\end{aligned}$$

by invariance of the trace distance under unitary transformation. Next, define $\boldsymbol{y}_W$ to be the outcome obtained by measuring $\rho_{\epsilon,WU_{\boldsymbol{x}}}^{\otimes n}$ with $\mathcal{M}$, and let $\boldsymbol{W} \in \mathsf{U}(d)$ be a Haar-random unitary operator chosen independently of $\boldsymbol{x}$. We claim that

$$\mathop{\mathbb{E}}_{\boldsymbol{W} \sim \mathrm{Haar}} I(\boldsymbol{x} : \boldsymbol{y}_{\boldsymbol{W}}) \leq I(\boldsymbol{U} : \boldsymbol{z}). \tag{9}$$

Let $p_{\boldsymbol{y}|W,x}$ to be the distribution of $\boldsymbol{y}_W$ given $\boldsymbol{x} = x$. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{W}} I(\boldsymbol{x} : \boldsymbol{y}_{\boldsymbol{W}}) &= \mathop{\mathbb{E}}_{\boldsymbol{W}} H\left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} p_{\boldsymbol{y}|W,\boldsymbol{x}} \right) - \mathop{\mathbb{E}}_{\boldsymbol{W}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} H(p_{\boldsymbol{y}|W,\boldsymbol{x}}) \\
&\leq H\left(\mathop{\mathbb{E}}_{\boldsymbol{W}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} p_{\boldsymbol{y}|W,\boldsymbol{x}} \right) - \mathop{\mathbb{E}}_{\boldsymbol{W}} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} H(p_{\boldsymbol{y}|W,\boldsymbol{x}}) \\
&= H\left(\mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} \mathop{\mathbb{E}}_{\boldsymbol{W}} p_{\boldsymbol{y}|\boldsymbol{W},\boldsymbol{x}} \right) - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim [N]} \mathop{\mathbb{E}}_{\boldsymbol{W}} H(p_{\boldsymbol{y}|\boldsymbol{W},\boldsymbol{x}}).
\end{aligned} \tag{10}$$

Where the first line follows from the definition of mutual information, the second line uses the concavity of entropy, and in the final line we make use of the independence of $\boldsymbol{x}$ and random unitary operator $\boldsymbol{W}$. Furthermore, by right-invariance of the Haar measure we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{W}} p_{\boldsymbol{y}|\boldsymbol{W},x} &= \mathop{\mathbb{E}}_{\boldsymbol{W}} \mathrm{diag}\left(\mathcal{M}(\rho_{\epsilon,\boldsymbol{W}U_x}^{\otimes n}) \right) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{W}} \mathrm{diag}\left(\mathcal{M}(\rho_{\epsilon,\boldsymbol{W}}^{\otimes n}) \right) \\
&= p_{\boldsymbol{z}}.
\end{aligned} \tag{11}$$

Similarly, we have for any $x \in [N]$ that

$$\mathop{\mathbb{E}}_{\boldsymbol{W}} H(p_{\boldsymbol{y}|\boldsymbol{W},x}) = \mathop{\mathbb{E}}_{\boldsymbol{U}} H(p_{\boldsymbol{z}}). \tag{12}$$

By substituting Eqs. (11) and (12) into Eq. (10) we arrive at the inequality in Eq. (9). We may once again invoke a probabilistic existence argument: since the expectation of $I(\boldsymbol{x} : \boldsymbol{y}_{\boldsymbol{W}})$ over unitary operators $\boldsymbol{W}$ is at most $I(\boldsymbol{U} : z)$, there exists at least one unitary operator $V \in \mathsf{U}(d)$ for which the inequality $I(\boldsymbol{x} : \boldsymbol{y}_V) \leq I(\boldsymbol{U} : z)$ holds. The proposition follows by considering the set of quantum states $\mathcal{S} := \{\rho_{\epsilon,VU_1}, \rho_{\epsilon,VU_2}, \ldots, \rho_{\epsilon,VU_N}\}$. □

Note that in this proposition the measurements performed on the product state can be arbitrary.

4.2 Lower bounds for nonadaptive measurements

In light of Proposition 4.3, in order to prove limitations of algorithms for tomography, it suffices to bound quantities of the form $I(\boldsymbol{U} : z)$ for Haar-random $\boldsymbol{U} \in \mathsf{U}(d)$ and measurement outcome z. To this end, it is helpful to establish the following relations based on Haar integration.

Lemma 4.4. *Fix an $\epsilon \in (0,1)$ and a positive integer $d > 1$. Let $\boldsymbol{U} \in \mathsf{U}(d)$ be a Haar-random unitary operator, $M \in \mathsf{Psd}(d)$ be a positive semidefinite operator such that $M \preceq \mathbb{1}$, $\rho_{\epsilon,U} \in \mathsf{D}(d)$ be defined as in Eq. (2) for each $U \in \mathsf{U}(d)$, and $w := \mathrm{Tr}(M)/d$. It holds that*

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathrm{Tr}\left(M\rho_{\epsilon,\boldsymbol{U}}\right) = w,$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \left(\mathrm{Tr}\left(M\rho_{\epsilon,\boldsymbol{U}}\right)\right)^2 \leq w^2\left(1 + \frac{\epsilon^2}{d+1} \cdot \min\left\{1, \frac{1}{w(d-1)}\right\}\right).$$

Proof. We defer the calculation of some Haar integrals to Appendix A. By the definition of $\rho_{\epsilon,U}$ in Eq. (2) the first expectation is

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}\left(M\rho_{\epsilon,\boldsymbol{U}}\right) = \frac{2\epsilon}{d} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}\left(M\boldsymbol{U}Q_{d/2}\boldsymbol{U}^\dagger\right) + (1-\epsilon)w.$$

Recall that $Q_{d/2} \in \mathsf{Psd}(d)$ is a rank-$d/2$ orthogonal projection operator. By Proposition A.2 in Appendix A and the linearity of trace we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}\left(M\boldsymbol{U}Q_{d/2}\boldsymbol{U}^\dagger\right) = \frac{\mathrm{Tr}(M)}{2}.$$

This leads to the first identity in the lemma. For the second expectation in the lemma, note that by substituting the definition of $\rho_{\epsilon,U}$ and expanding we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \left(\mathrm{Tr}(M\rho_{\epsilon,\boldsymbol{U}})\right)^2 &= \frac{4\epsilon^2}{d^2} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \left(\mathrm{Tr}(M\boldsymbol{U}Q_{d/2}\boldsymbol{U}^\dagger)\right)^2 + w^2(1-\epsilon^2) \\ &= \frac{4\epsilon^2}{d^2} \mathrm{Tr}\left(M^{\otimes 2} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} (\boldsymbol{U}Q_{d/2}\boldsymbol{U}^\dagger)^{\otimes 2}\right) + w^2(1-\epsilon^2). \qquad (13)\end{aligned}$$

The Haar integral on the right-hand side is evaluated explicitly in Proposition A.3 by setting the rank parameters to $r_1 = r_2 = d/2$. This yields

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} (\boldsymbol{U}Q_{d/2}\boldsymbol{U}^\dagger)^{\otimes 2} = \frac{1}{4(d^2-1)}\left[(d^2-2)\mathbb{1} + dW\right]$$

where the identity and swap operator W act on $(\mathbb{C}^d)^{\otimes 2}$. Substituting into Eq. (13) and making use of the identity $\mathrm{Tr}(W(A \otimes B)) = \mathrm{Tr}(AB)$ we find that the right-hand side is equal to

$$\frac{\epsilon^2(d^2-2)(\mathrm{Tr}(M))^2}{d^2(d^2-1)} + \frac{\epsilon^2 \mathrm{Tr}(M^2)}{d(d^2-1)} + w^2(1-\epsilon^2) = \frac{\epsilon^2(d^2-2)w^2}{d^2-1} + \frac{\epsilon^2 \mathrm{Tr}(M^2)}{d(d^2-1)} + w^2(1-\epsilon^2)$$
$$= w^2 + \frac{\epsilon^2(\mathrm{Tr}(M^2) - dw^2)}{d(d^2-1)}. \tag{14}$$

Assume for now that

$$\mathrm{Tr}(M^2) - dw^2 \leq \min\{w^2 d(d-1), wd\}. \tag{15}$$

Then the right-hand side of Eq. (14) is at most

$$w^2 + \frac{\epsilon^2}{d(d^2-1)} \cdot \min\{w^2 d(d-1), wd\} = w^2 \left(1 + \frac{\epsilon^2}{d+1} \cdot \min\left\{1, \frac{1}{w(d-1)}\right\}\right)$$

as required. To prove the inequality in Eq. (15), we make use of the relations $\mathrm{Tr}(M^2) \leq (\mathrm{Tr}(M))^2 = w^2 d^2$ and $\mathrm{Tr}(M^2) \leq \mathrm{Tr}(M) = wd$ both of which follow from the property that $0 \preceq M \preceq \mathbb{1}$. The second bound of wd follows from the nonnegativity of dw^2. □

This leads us to the lower bounds stated in Theorem 4.5 below. Intuitively, the theorem establishes the following property: for the family of quantum states of the form in Eq. (2), the ability to distinguish the distribution over outcomes of a measurement from some fixed distribution—as quantified by their χ^2-divergence—is small on average, no matter the measurement performed. In proving this theorem, our analysis is simplified due to Lemma 4.1 as well as techniques for Haar integration based on permutation invariance. (We refer the interested reader to Section 7.2 of Ref. [Wat18] for more on this topic.) We also do not assume that the measurement operators which comprise a given POVM are rank-one, as has been considered in other works [HHJ+17, HKP21, CCHL22]. This allows us to conclude the novel $\Omega(d^4/\epsilon^2)$ lower bound in the constant-outcome case, in addition to laying the groundwork for the results in Sections 5 and 6.

Theorem 4.5. *Fix an $\epsilon \in (0,1)$ and a positive integer $d > 1$. Let $\rho_{\epsilon,U} \in \mathsf{D}(d)$ be defined as in Eq. (2), $\boldsymbol{U} \in \mathsf{U}(d)$ be a Haar-random unitary operator, and $\boldsymbol{z}$ be the outcome of a measurement $\mathcal{M} \in \Xi(d)$ performed on the random state $\rho_{\epsilon,\boldsymbol{U}}$ such that $p_{\boldsymbol{z}|U} = \mathrm{diag}(\mathcal{M}(\rho_{\epsilon,U}))$ for every $U \in \mathsf{U}(d)$. Then*

$$\mathop{\mathbb{E}}_{\boldsymbol{U} \sim \mathrm{Haar}} \mathrm{D}_{\chi^2}(p_{\boldsymbol{z}|\boldsymbol{U}} \,\|\, p_{\boldsymbol{z}}) \leq \frac{\epsilon^2}{d+1} \cdot \min\left\{1, \frac{\mathrm{rank}(\mathcal{M})}{d-1}\right\}.$$

Proof. Let $\mathcal{Z}$ be an alphabet denoting the set of possible outcomes of the measurement $\mathcal{M}$, such that $z \in \mathcal{Z}$ if and only if $|z\rangle\langle z| \in \mathrm{im}(\mathcal{M})$ for orthonormal $\{|z\rangle\}$. By Definition 2.8 we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U} \sim \mathrm{Haar}} \mathrm{D}_{\chi^2}(p_{\boldsymbol{z}|\boldsymbol{U}} \,\|\, p_{\boldsymbol{z}}) = \sum_{z \in \mathcal{Z}} \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \mathrm{Haar}} \frac{p_{\boldsymbol{z}|\boldsymbol{U}}(z)^2}{p_{\boldsymbol{z}}(z)} - 1 \tag{16}$$

where for fixed $U \in \mathsf{U}(d)$ the conditional probabilities may be written as $p_{\boldsymbol{z}|U}(z) = \mathrm{Tr}(M_z \rho_{\epsilon,U})$ for the POVM $(M_z)_z$ corresponding to the measurement $\mathcal{M}$, and the marginal probabilities in the denominator can be written as $p_{\boldsymbol{z}}(z) = \mathbb{E}_{\boldsymbol{U} \sim \mathrm{Haar}} \mathrm{Tr}(M_z \rho_{\epsilon,\boldsymbol{U}})$. Let $w(z) = \mathrm{Tr}(M_z)/d$ for all $z \in \mathcal{Z}$. By Lemma 4.4 the right-hand side of Eq. (16) is at most

$$\sum_{z \in \mathcal{Z}} w(z) \left(1 + \frac{\epsilon^2}{d+1} \cdot \min\left\{1, \frac{1}{w(z)(d-1)}\right\}\right) - 1 = \frac{\epsilon^2}{d+1} \cdot \min\left\{1, \frac{|\mathcal{Z}|}{d-1}\right\}. \tag{17}$$

Since $|\mathcal{Z}| = \mathrm{rank}(\mathcal{M})$, this concludes the proof. □

In the above theorem, the rank of $\mathcal{M}$ may be interpreted as the maximum number of outcomes that can be resolved using the measurements, under the assumption that the learner discards each copy of the state after measuring it.

We now have the tools we need to prove the two lower bounds for the nonadaptive case shown in Table 1. The first is a result originally due to Ref. [HHJ+17].

Corollary 4.6 (Special case of Theorem 4 in Ref. [HHJ+17])**.** *Let $\epsilon \in (0,1)$. Any procedure for quantum tomography of d-dimensional quantum states that is $\epsilon/4$-accurate in trace distance using nonadaptive, single-copy measurements requires $n \in \Omega\left(d^3/\epsilon^2\right)$ samples of the unknown state.*

Proof. Let $\mathcal{M} = \mathcal{M}_1 \otimes \cdots \otimes \mathcal{M}_n \in \Xi(d^n)$ be the single-copy, nonadaptive measurement which is performed on the n copies of the unknown state to do tomography. By Proposition 4.3, there exists an $\epsilon/2$-packing $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ of $N \in \exp(\Omega(d^2))$ quantum states of the form in Lemma 3.3 such that the following holds. Let $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the outcome of the measurement $\mathcal{M}$ when performed on n copies of the random state $\rho_{\boldsymbol{x}}$. Then $I(\boldsymbol{x}:\boldsymbol{y}) \leq I(\boldsymbol{U}:\boldsymbol{z})$ where $\boldsymbol{U}$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are defined as in the proposition: $\boldsymbol{U} \in \mathsf{U}(d)$ is Haar-random, and $\boldsymbol{z}_k$ is the measurement outcome obtained by measuring $\rho_{\epsilon,\boldsymbol{U}}$ with $\mathcal{M}_k$, for each $k \in [n]$. Since the random variables $\boldsymbol{z}_k$ are independent given $\boldsymbol{U}$, using the chain rule for mutual information, and monotonicity of entropy under conditioning, we have

$$\begin{aligned} I(\boldsymbol{U}:\boldsymbol{z}) &= \sum_{k=1}^{n} H(\boldsymbol{z}_k|\boldsymbol{z}_{<k}) - H(\boldsymbol{z}_k|\boldsymbol{z}_{<k},\boldsymbol{U}) \\ &= \sum_{k=1}^{n} H(\boldsymbol{z}_k|\boldsymbol{z}_{<k}) - H(\boldsymbol{z}_k|\boldsymbol{U}) \\ &\leq \sum_{k=1}^{n} H(\boldsymbol{z}_k) - H(\boldsymbol{z}_k|\boldsymbol{U}) \\ &= \sum_{k=1}^{n} I(\boldsymbol{U}:\boldsymbol{z}_k). \end{aligned}$$

We apply Corollary 4.2 to bound mutual information from above in terms of χ^2-divergence, and then Theorem 4.5 to each of the terms in this sum to get

$$\begin{aligned} I(\boldsymbol{x}:\boldsymbol{y}) &\leq \frac{n}{\ln(2)} \left(\max_{k\in[n]} \mathop{\mathbb{E}}_{\boldsymbol{u}\sim\mathrm{Haar}} \mathsf{D}_{\chi^2}(p_{\boldsymbol{z}_k|\boldsymbol{u}} \,\|\, p_{\boldsymbol{z}_k}) \right) \\ &\leq \frac{n\epsilon^2}{\ln(2)(d+1)} \min\left\{ 1, \frac{\max_{k\in[n]} \mathrm{rank}(\mathcal{M}_k)}{d-1} \right\} \qquad (18) \\ &\leq \frac{n\epsilon^2}{\ln(2)(d+1)}. \qquad (19) \end{aligned}$$

Under the assumption that the tomography algorithm gives us a state that is accurate to within $\epsilon/4$ in trace distance, the measurement $\mathcal{M}$ can be used to decode $\boldsymbol{x}$ with some constant probability of success. By Fano's inequality as well as the bound in Eq. (19), it holds that

$$\frac{n\epsilon^2}{\ln(2)(d+1)} \in \Omega(d^2)$$

which is true if and only if $n \in \Omega(d^3/\epsilon^2)$. □

Corollary 4.7. *Any procedure for quantum tomography of d-dimensional quantum states that is ϵ-accurate in trace distance using nonadaptive, single-copy measurements, each with at most ℓ outcomes, requires $n \in \Omega\left(d^4/\epsilon^2\ell\right)$ samples of the unknown state.*

Proof. The proof is identical to that for Corollary 4.6 except we use Theorem 4.5 to bound the right-hand side of Eq. (18) in terms of the maximum rank of the measurement operators, which in this case is at most ℓ by assumption. We then have

$$\frac{n\epsilon^2\ell}{\ln(2)(d^2-1)} \in \Omega(d^2)$$

which is true if and only if $n \in \Omega(d^4/\epsilon^2\ell)$. □

Corollary 4.7 implies that there is a strong sense in which the folklore "Pauli tomography" algorithm—which has an upper bound of $O(d^4/\epsilon^2)$ measurements—is sample-optimal: amongst all possible strategies making use of constant-outcome (and in particular, two-outcome) measurements, there is no way to perform tomography that is more efficient. Note that here it is assumed that each copy of the state is discarded upon performing the measurement. In the more general case where one may perform further non-adaptive measurements on post-measurement states, the lower bound from Corollary 4.6 applies.

4.3 Rank-dependent bounds

In this section we derive lower bounds for state tomography using non-adaptive single-copy measurements, when the states are known to have bounded rank.

We consider a different packing of states defined as follows (cf. Ref. [HHJ+17, Section VI.B]). Fix $\nu \in (0,1)$, positive integers $d \geq 3$ and $r \in [1, d/3]$. For $i \in [r]$, define the pure state

$$|\psi_{\nu,i}\rangle := \sqrt{1-\nu}\,|d+1-i\rangle + \sqrt{\nu}\,|i\rangle \ . \tag{20}$$

For a unitary operator $U \in \mathsf{U}(\mathbb{C}^{d-r})$, which we extend to $\mathbb{C}^d$ by taking a direct sum with the identity, define the rank r state

$$\sigma_{\nu,U} := U\left(\frac{1}{r}\sum_{i=1}^{r}|\psi_{\nu,i}\rangle\langle\psi_{\nu,i}|\right)U^\dagger \ . \tag{21}$$

There is a large packing of states of this form.

Lemma 4.8 (part of Lemma 7 in Ref. [HHJ+17])**.** *For any $\nu \in (0,1/4)$ there exists a $\sqrt{\nu}/4$-packing $\mathcal{S} \subset \mathsf{D}(d)$ of N quantum states of the form in Eq. (21), with $N \in \exp\left(\Omega(rd)\right)$.*

By the same reasoning as for Proposition 4.3, we have

Lemma 4.9. *Let $\boldsymbol{U}$ be a Haar-random unitary operator over $\mathbb{C}^{d-r}$ and $\boldsymbol{z}$ be the outcome obtained upon measuring the random state $\sigma_{\nu,\boldsymbol{U}}^{\otimes n}$ with some measurement $\mathcal{M}$. There exists a set of N quantum states $\mathcal{S} := \{\sigma_1,\ldots,\sigma_N\} \subset \mathsf{D}(d)$ with $N \in \exp(\Omega(rd))$ of the form in Eq. (21) which is a $\sqrt{\nu}/4$-packing and satisfies $I(\boldsymbol{x}:\boldsymbol{y}) \leq I(\boldsymbol{U}:\boldsymbol{z})$, where $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and $\boldsymbol{y}$ is the outcome obtained from measuring the random state $\sigma_{\boldsymbol{x}}^{\otimes n}$ with $\mathcal{M}$.*

We bound some measurement statistics associated with a random state of the form in Eq. (21) in preparation for the main results of this section. Let $\Gamma_1 := \sum_{i=1}^{d-r}|i\rangle\langle i|$ and $\Gamma_0 := \mathbb{1} - \Gamma_1$.

Lemma 4.10. *Let $\boldsymbol{U}$ be a Haar-random unitary operator on $\mathbb{C}^{d-r}$, $M \in \mathsf{Psd}(d)$ be a positive semidefinite operator such that $M \preceq \mathbb{1}$, $\sigma_{\nu,U} \in \mathsf{D}(d)$ be defined as in Eq. (21) for each unitary operator U on $\mathbb{C}^{d-r}$, and*

$$w \coloneqq \frac{(1-\nu)}{r}\operatorname{Tr}(M\Gamma_0) + \frac{\nu}{d-r}\operatorname{Tr}(M\Gamma_1) \ . \tag{22}$$

Then

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \operatorname{Tr}\left(M\sigma_{\nu,\boldsymbol{U}}\right) = w \ ,$$

and

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{U}} \left(\operatorname{Tr}\left(M\sigma_{\nu,\boldsymbol{U}}\right)\right)^2 \leq w^2 &+ \frac{2\nu^2}{(d-r)^4}\left(\operatorname{Tr}(M\Gamma_1)\right)^2 + \frac{3\nu^2}{r(d-r)^2}\operatorname{Tr}\left((M\Gamma_1)^2\right) \\ &+ \frac{2\nu(1-\nu)}{r^2(d-r)}\operatorname{Tr}\left(M\Gamma_1 M\Gamma_0\right) \ .\end{aligned}$$

Proof. Due to the ± 1 symmetry of the Haar measure, the terms with an odd number of occurrences of $\boldsymbol{U}$ or $\boldsymbol{U}^\dagger$ in the expansion of $\sigma_{\nu,\boldsymbol{U}}$ and $\sigma_{\nu,\boldsymbol{U}}^{\otimes 2}$ evaluate to 0 in expectation. The expectation of $\operatorname{Tr}(M\sigma_{\nu,\boldsymbol{U}})$ then follows as before. For the bound on the second expectation, note that

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \left(\operatorname{Tr}\left(M\sigma_{\nu,\boldsymbol{U}}\right)\right)^2 = \mathop{\mathbb{E}}_{\boldsymbol{U}} \operatorname{Tr}\left((M\otimes M)\left(\sigma_{\nu,\boldsymbol{U}}\otimes\sigma_{\nu,\boldsymbol{U}}\right)\right) \ .$$

Define $\widetilde{\Gamma}_1 \coloneqq \sum_{i=1}^r |i\rangle\langle i|$, $\widetilde{\Gamma}_{10} \coloneqq \sum_{i=1}^r |i\rangle\langle d+1-i|$, and $\widetilde{\Gamma}_{01} \coloneqq \sum_{i=1}^r |d+1-i\rangle\langle i|$. Combining the ± 1 symmetry of the Haar measure with Propositions A.4 and A.3, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{U}} \sigma_{\nu,\boldsymbol{U}}^{\otimes 2} &= \mathop{\mathbb{E}}_{\boldsymbol{U}}\Big[\frac{(1-\nu)^2}{r^2}\Gamma_0^{\otimes 2} + \frac{\nu(1-\nu)}{r^2}\Big(\Gamma_0\otimes \boldsymbol{U}\widetilde{\Gamma}_1\boldsymbol{U}^\dagger \\ &\qquad + \boldsymbol{U}\widetilde{\Gamma}_1\boldsymbol{U}^\dagger\otimes\Gamma_0 + \left(\boldsymbol{U}(\widetilde{\Gamma}_{10}+\widetilde{\Gamma}_{01})\boldsymbol{U}^\dagger\right)^{\otimes 2}\Big) + \frac{\nu^2}{r^2}\left(\boldsymbol{U}\widetilde{\Gamma}_1\boldsymbol{U}^\dagger\right)^{\otimes 2}\Big] \\ &= \frac{(1-\nu)^2}{r^2}\Gamma_0^{\otimes 2} + \frac{\nu(1-\nu)}{r(d-r)}\Big(\Gamma_0\otimes\Gamma_1 + \Gamma_1\otimes\Gamma_0\Big) \\ &\quad + \frac{\nu(1-\nu)}{r^2(d-r)}\sum_{i=1}^{r}\sum_{k=1}^{d-r}\left(|k\rangle\langle d+1-i|\otimes|d+1-i\rangle\langle k| + |d+1-i\rangle\langle k|\otimes|k\rangle\langle d+1-i|\right) \\ &\quad + \frac{\nu^2}{r(d-r)((d-r)^2-1)}\Big((r(d-r)-1)\mathbb{1} + (d-2r)W\Big)(\Gamma_1\otimes\Gamma_1) \ ,\end{aligned}$$

where $\mathbb{1}$ and W are the identity and swap operators on $\mathbb{C}^d\otimes\mathbb{C}^d$, respectively. We have

$$\frac{1}{(d-r)^2-1} \leq \frac{1}{(d-r)^2}\left(1+\frac{2}{(d-r)^2}\right) \ ,$$

since $(d-r)^2 \geq 2$. Noting that $\operatorname{Tr}((A\otimes B)W) = \operatorname{Tr}(AB)$ and $d-2r \leq d-r$, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{U}} \left(\operatorname{Tr}\left(M\sigma_{\nu,\boldsymbol{U}}\right)\right)^2 \leq \frac{(1-\nu)^2}{r^2}\left(\operatorname{Tr}(M\Gamma_0)\right)^2 &+ \frac{2\nu(1-\nu)}{r(d-r)}\operatorname{Tr}(M\Gamma_0)\operatorname{Tr}(M\Gamma_1) \\ &+ \frac{2\nu(1-\nu)}{r^2(d-r)}\operatorname{Tr}(M\Gamma_1 M\Gamma_0) \\ &+ \frac{\nu^2}{(d-r)^2}\left(1+\frac{2}{(d-r)^2}\right)\left(\operatorname{Tr}(M\Gamma_1)\right)^2 + \frac{3\nu^2}{r(d-r)^2}\operatorname{Tr}\left((M\Gamma_1)^2\right) \ .\end{aligned}$$

The bound in the statement of the lemma now follows by the definition of w in Eq. (22). □

We now prove a slightly stronger lower bound for the tomography of states with bounded rank as compared with the bound implied by Ref. [HHJ+17]; see the remark following Theorem 4 in this reference. The proof of the said theorem assumes that the rank of the input states is strictly greater than one, so the bound for pure states in the theorem we establish below appears to be new.

Recall that $d \geq 3$ and $r \in [1, d/3]$.

Theorem 4.11. *Let $\epsilon \in (0, 1/8)$. Any algorithm for quantum tomography of rank r quantum states in d-dimensions that uses non-adaptive, single-copy measurements and produces an approximation within ϵ in trace distance with positive constant probability requires $\Omega\left(r^2 d/\epsilon^2\right)$ samples of the unknown state.*

Proof. We proceed as in the proof of Corollary 4.6. Consider any algorithm as in the statement of the theorem, and let $\mathcal{M} := \mathcal{M}_1 \otimes \cdots \otimes \mathcal{M}_n \in \Xi(d^n)$ be the single-copy, non-adaptive measurement performed by it on the n copies of the unknown pure state.

We take $\nu := 64\epsilon^2$. By Lemma 4.9, there exists a 2ϵ-packing $\mathcal{S} := \{\sigma_1, \ldots, \sigma_N\} \subset \mathsf{D}(d)$ of quantum states of the form in Eq. (21), with $N \in \exp(\Omega(rd))$, which satisfy the following property. Let $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and $\boldsymbol{y} := (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the outcome of the measurement $\mathcal{M}$ when performed on n copies of the random state $\sigma_{\boldsymbol{x}}$. Then $I(\boldsymbol{x} : \boldsymbol{y}) \leq I(\boldsymbol{U} : \boldsymbol{z})$, where $\boldsymbol{U}$ and $\boldsymbol{z} := (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are defined as in the lemma: $\boldsymbol{U}$ is a Haar-random unitary operator over $\mathbb{C}^{d-r}$, and $\boldsymbol{z}_k$ is the measurement outcome obtained by measuring $\sigma_{\nu,\boldsymbol{U}}$ with $\mathcal{M}_k$, for each $k \in [n]$.

As in Corollary 4.6, using Lemma 4.10, we get

$$\begin{aligned} I(\boldsymbol{x} : \boldsymbol{y}) &\leq \frac{n}{\ln(2)} \left(\max_{k \in [n]} \mathop{\mathbb{E}}_{\boldsymbol{U}} \mathsf{D}_{\chi^2}(p_{\boldsymbol{z}_k|\boldsymbol{U}} \parallel p_{\boldsymbol{z}_k}) \right) \\ &\leq \frac{n}{\ln(2)} \left[\frac{2\nu^2}{(d-r)^4} \sum_z \frac{1}{w_z} \left(\mathrm{Tr}(M_z \Gamma_1)\right)^2 + \frac{3\nu^2}{r(d-r)^2} \sum_z \frac{1}{w_z} \mathrm{Tr}\left((M_z \Gamma_1)^2\right) \right. \\ &\qquad \left. + \frac{2\nu(1-\nu)}{r^2(d-r)} \sum_z \frac{1}{w_z} \mathrm{Tr}\left(M_z \Gamma_1 M_z \Gamma_0\right) \right] , \end{aligned} \tag{23}$$

where (M_z) is one of the n measurements $\mathcal{M}_k$ which maximizes the expected χ^2-divergence above, and w_z is given by Eq. (22) with $M := M_z$.

Since the algorithm approximates the unknown state to within ϵ and the states σ_x form a 2ϵ-packing, the algorithm correctly identifies $\boldsymbol{x}$ with positive constant probability. By Fano's Inequality, we have

$$I(\boldsymbol{x} : \boldsymbol{y}) \in \Omega(rd) \ . \tag{24}$$

To conclude the lower bound of $\Omega(r^2 d/\epsilon^2)$ on n, it suffices to show that the right side of Eq. (23) is of the order of $n\nu/r$.

We have $\mathrm{Tr}\left((M_z \Gamma_1)^2\right) \leq \left(\mathrm{Tr}(M_z \Gamma_1)\right)^2$, and

$$\begin{aligned} \mathrm{Tr}\left(M_z \Gamma_1 M_z \Gamma_0\right) &\leq \left(\mathrm{Tr}(\Gamma_1 M_z^2 \Gamma_1)\right)^{1/2} \left(\mathrm{Tr}(\Gamma_0 M_z^2 \Gamma_0)\right)^{1/2} \\ &\leq \left(\mathrm{Tr}(M_z \Gamma_1)\right) \left(\mathrm{Tr}(M_z \Gamma_0)\right) \ . \end{aligned}$$

Combining this with $2r/(d-r)^2 \leq 1$ and the definition of w_z (in particular that $w_z \geq (\nu/(d-r)) \mathrm{Tr}(M_z \Gamma_1)$), we get the desired bound on the right hand side of Eq. (23):

$$\frac{n}{\ln(2)} \cdot \frac{\nu}{r(d-r)} \sum_z \frac{\mathrm{Tr}(M_z \Gamma_1)}{w_z} \left[\frac{4\nu}{(d-r)} \mathrm{Tr}(M_z \Gamma_1) + \frac{2(1-\nu)}{r} \mathrm{Tr}(M_z \Gamma_0) \right]$$

$$\leq \frac{n}{\ln(2)} \cdot \frac{4\nu}{r(d-r)} \sum_z \mathrm{Tr}(M_z \Gamma_1)$$
$$\leq \frac{4n\nu}{r\ln(2)} \ .$$

This completes the proof. □

The same approach allows us to derive stronger bounds when the measurements used by the tomography algorithm have a constant number of outcomes. Again, $d \geq 3$ and $r \in [1, d/3]$.

Theorem 4.12. *Let $\epsilon \in (0, 1/8)$. Any algorithm for quantum tomography of rank r quantum states in d-dimensions that uses non-adaptive, single-copy measurements* with at most ℓ outcomes *and produces an approximation within ϵ in trace distance with positive constant probability requires $\Omega\left(r^2 d^2 / \ell \epsilon^2\right)$ samples of the unknown state.*

Proof. The proof largely proceeds as for Theorem 4.11. We use the same notation here, and only indicate where we deviate from that proof.

To conclude the claimed lower bound on n, it suffices to show that the right side of Eq. (23) is of the order of $n\ell\nu/rd$. We have

$$\mathrm{Tr}\left((M_z \Gamma_1)^2\right) = \mathrm{Tr}\left((\Gamma_1 M_z \Gamma_1)^2\right) \leq \mathrm{Tr}(\Gamma_1 M_z \Gamma_1) \ ,$$

and

$$\mathrm{Tr}(M_z \Gamma_1 M_z \Gamma_0) \leq \mathrm{Tr}(M_z^2 \Gamma_0) \leq \mathrm{Tr}(M_z \Gamma_0) \ ,$$

since $\Gamma_1 \preceq \mathbb{1}$, $M_z^2 \preceq M_z$, and $\Gamma_0 \succeq 0$. Further observe that $w_z \geq \nu \,\mathrm{Tr}(M_z \Gamma_1)/(d-r)$, $2r/(d-r)^2 \leq 1$, and $\mathrm{Tr}(M_z \Gamma_1) \leq d - r$. We thus get the following bound on the right side of Eq. (23):

$$\begin{aligned}
&\frac{n}{\ln(2)} \cdot \frac{\nu}{r(d-r)} \left[\sum_z \frac{1}{w_z} \left(\frac{3\nu}{d-r} \mathrm{Tr}(M_z \Gamma_1) + \frac{2(1-\nu)}{r} \mathrm{Tr}(M_z \Gamma_0) \right) \right. \\
&\qquad \left. + \frac{2r}{(d-r)^2} \sum_z \frac{\nu}{w_z (d-r)} \left(\mathrm{Tr}(M_z \Gamma_1)\right)^2 \right] \\
&\leq \frac{n}{\ln(2)} \cdot \frac{\nu}{r(d-r)} \left[3\ell + \frac{2r}{(d-r)^2} \sum_z \mathrm{Tr}(M_z \Gamma_1) \right] \\
&\leq \frac{n}{\ln(2)} \cdot \frac{4\nu\ell}{r(d-r)} \ .
\end{aligned}$$

Here, we bounded the term in the curved parentheses in the first line by $3w_z$. The result is the desired bound on the right hand side of Eq. (23). □

5 Lower bounds for tomography with adaptive measurements

In Section 4 we saw that it is possible to derive lower bounds on tomography that are stronger than the bound of $\Omega(d^2)$ obtained by a direct application of Holevo's theorem, by considering restricted measurements. (See also Theorem 4 in Ref. [HHJ+17], or Chapter 5 in Wright's Ph.D. thesis [Wri16].) Specifically, we were able to show optimal lower bounds on tomography in the nonadaptive case for both constant-outcome and arbitrary measurements. In this section we consider a different kind of restriction on the measurements; namely, the measurements may be adaptively chosen, so long as they are chosen from a finite set of m different measurements.

We show that even when we have a choice of $\exp(d)$ different measurements, the $\Omega(d^3/\epsilon^2)$ lower bound from the previous section continues to hold. In particular, this lower bound applies to the case when the single-copy measurements are all efficiently implementable, i.e., implementable as uniformly generated polylog(d)-size quantum circuits. In other words, adaptivity does not help while using measurements involving quantum computation with a number of gates growing only polynomially in the number of qubits measured. In Section C we explain how these results can also be applied to rule out an advantage for adaptivity using a possibly infinite number of measurement settings (i.e., when the measurements are chosen from an infinite set), whenever the measurements are implementable with polylog(d)-size quantum circuits.

The approach we take also leads to a lower bound in the adaptive, constant-outcome case which generalizes an earlier result due to Ref. [FGLE12]. There it is shown that $\Omega(d^4/\log(d))$ copies of the state are required when one is restricted to two-outcome, projective, possibly adaptive Pauli measurements.

5.1 Distinguishability of a hard ensemble

To arrive at lower bounds robust to adaptivity, we once again appeal to difficult instances of the quantum state discrimination problem. This time, however, we construct a packing of quantum states with the additional requirement that all selected states lead to uninformative measurements using *any* measurement from a fixed set of possibilities. Such a construction is enabled by a tail bound on the χ^2-divergence quantities we have been considering, so that most states, in addition to being well-separated from previous choices, offer only uninformative measurement statistics. Similar tail bounds have been derived in prior work for the purpose of showing unconditional lower bounds for quantum state certification with adaptive measurements [BCL20].

The concentration of measure property we invoke to arrive at our tail bounds follow from log-Sobolev inequalities, and is analogous to Lévy's Lemma for functions on the unit sphere [Mec19]. A detailed discussion is beyond the scope of this work, but roughly speaking these imply that sufficiently well-behaved functions of unitary operators concentrate strongly around their expectation. In particular, we have the following theorem.

Theorem 5.1 (Special case of Theorem 5.17 in Ref. [Mec19])**.** *Let $d > 1$ be a positive integer, $f : \mathsf{U}(d) \to \mathbb{R}$ be κ-Lipschitz with respect to the metric induced by the Frobenius norm, and let $\mu := \mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} f(\boldsymbol{U})$. Then, for any $t > 0$, it holds that*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}[f(\boldsymbol{U}) \geq \mu + t] \leq \exp\left(-\frac{(d-2)t^2}{24\kappa^2}\right).$$

Before proceeding, we introduce a more convenient short-hand notation for the χ^2-divergence quantities which arose in the analysis in the previous section.

Definition 5.2. For any $\epsilon \in (0,1)$ and positive integer $d > 1$, let $\rho_{\epsilon,U} \in \mathsf{D}(d)$ be defined as in Eq. (2). We define the function $F^{\chi^2}_{\epsilon,d} : \Xi(d) \times \mathsf{U}(d) \to \mathbb{R}$ by

$$F^{\chi^2}_{\epsilon,d}(\mathcal{M}, U) := \mathrm{D}_{\chi^2}(p_{\boldsymbol{z}|U} \,\|\, w)$$

for all $U \in \mathsf{U}(d)$ and $\mathcal{M} \in \Xi(d)$, where $p_{\boldsymbol{z}|U} := \mathrm{diag}(\mathcal{M}(\rho_{\epsilon,U}))$ and $w := \mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} p_{\boldsymbol{z}|\boldsymbol{U}}$.

For any $\mathcal{M} \in \Xi(d)$ with corresponding measurement operators $\{M_z : z \in \mathcal{Z}\} \subset \mathsf{Psd}(d)$ and $U \in \mathsf{U}(d)$,

$$|F^{\chi^2}_{\epsilon,d}(\mathcal{M}, U)| = \sum_{z\in\mathcal{Z}} \frac{\mathrm{Tr}(M_z\rho_{\epsilon,U})^2}{w(z)} - 1 = \sum_{z\in\mathcal{Z}} \frac{\epsilon^2(2\,\mathrm{Tr}(M_z U Q_{d/2} U^\dagger)/d - w(z))^2}{w(z)},$$

where $w(z) := \mathrm{Tr}(M_z)/d$. Here, we have used the definition of $\rho_{\epsilon,U}$ from Eq. (2) to write the expression in terms of the operator $Q_{d/2}$. Since $0 \leq \mathrm{Tr}(M_z U Q_{d/2} U^\dagger)/d \leq 2w(z)$, we have $\|F^{\chi^2}_{\epsilon,d}\|_\infty \leq \epsilon^2$.

We turn to the tail bound which we use in this section to derive lower bounds in the case of adaptive measurements.

Lemma 5.3 (χ^2-squared tail bound)**.** *Fix an $\epsilon \in (0,1)$ and a positive integer $d \geq 4$. For any finite-outcome measurement $\mathcal{M} \in \Xi(d)$ it holds that*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U}) > \alpha + t\right] \leq \exp\left(-\frac{Cd^2t}{\epsilon^2}\right) \tag{25}$$

where $\alpha := c\epsilon^2/d$ and c, C are universal constants that we may take to be 2 and $1/(3\cdot 2^8)$, respectively. Furthermore, if $\mathcal{M}$ is restricted to having ℓ outcomes then the inequality holds with $\alpha := 4\ell\epsilon^2/3d^2$.

Proof. We first consider the case where the measurement $\mathcal{M}$ may have an arbitrary number of outcomes. Our goal is to prove that the random variable $F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U}) - c\epsilon^2/d$ is subexponential, where $\boldsymbol{U}$ is Haar-random. To accomplish this, we follow the approach in the proof of Lemma 7.6 in Ref. [BCL20]. Instead of bounding the tail of the random variable directly using Lemma 5.3, we consider its square root. We are then able to show a comparatively stronger bound on the Lipschitz constant of this function. Translating the resulting subgaussian tail on $\sqrt{F^{\chi^2}_{\epsilon,d}}$ into a subexponential tail on $F^{\chi^2}_{\epsilon,d}$ controls its deviations in the regime we care about. In particular, it suffices to show that the function f which acts on $U \in \mathsf{U}(d)$ as

$$f : U \mapsto \sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M},U)} - \mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} \sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{V})}$$

has a tail like $\exp(-\Omega(d^2t^2/\epsilon^2))$, for U selected randomly from the Haar distribution.

In more detail, note that

$$\mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} \sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{V})} \leq \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{V})} \leq \frac{\epsilon}{\sqrt{d}}$$

by the Jensen inequality and Theorem 4.5. Furthermore, the inequality

$$c\epsilon^2/d + t \geq \left(\sqrt{c\epsilon^2/d} + \sqrt{t}\right)^2/2$$

for any $t \geq 0$ entails that

$$\begin{aligned}\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U}) > \frac{c\epsilon^2}{d} + t\right] &= \Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U})} > \sqrt{\frac{c\epsilon^2}{d} + t}\right] \\ &\leq \Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U})} > \epsilon\sqrt{\frac{c}{2d}} + \sqrt{\frac{t}{2}}\right].\end{aligned}$$

By choosing $c := 2$ we find that if f has a tail of $\exp(-\Omega(d^2t^2/\epsilon^2))$ we get

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[F^{\chi^2}_{\epsilon,d}(\mathcal{M},\boldsymbol{U}) > \frac{c\epsilon^2}{d} + t\right] \leq \exp\left(-\frac{Cd^2t}{\epsilon^2}\right)$$

for some universal constant C, as required.

To arrive at the desired concentration of measure for f we invoke Theorem 5.1, according to which it suffices to show that f is $O(\epsilon/\sqrt{d})$-Lipschitz. Let $\boldsymbol{z}$ be the outcome obtained by measuring $\rho_{\epsilon,\boldsymbol{U}}$ with $\mathcal{M}$. It has conditional distribution $p_{\boldsymbol{z}|U}$ given $\boldsymbol{U} = U$. Also recall the distribution w over outcomes given by $w := \mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} p_{\boldsymbol{z}|\boldsymbol{U}}$. For arbitrary $U, V \in \mathsf{U}(d)$, by Definitions 5.2 and 2.8 and the Triangle Inequality, we have

$$\begin{aligned}
|f(U) - f(V)| &= \left| \sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M}, U)} - \sqrt{F^{\chi^2}_{\epsilon,d}(\mathcal{M}, V)} \right| \\
&= \left| \sqrt{\mathop{\mathbb{E}}_{z'\sim w} \left(\frac{p_{z|U}(z')}{w(z')} - 1 \right)^2} - \sqrt{\mathop{\mathbb{E}}_{z'\sim w} \left(\frac{p_{z|V}(z')}{w(z')} - 1 \right)^2} \right| \\
&\leq \sqrt{\mathop{\mathbb{E}}_{z'\sim w} \left(\frac{p_{z|U}(z')}{w(z')} - \frac{p_{z|V}(z')}{w(z')} \right)^2}.
\end{aligned}$$

Let $\{M_z : z \in \mathcal{Z}\}$ be the measurement operators corresponding to $\mathcal{M}$. We have that $p_{\boldsymbol{z}|U}(z) = \mathrm{Tr}(M_z \rho_{\epsilon,U})$ and $w(z) = \mathrm{Tr}(M_z)/d$ from Lemma 4.4. Recalling the definition of $\rho_{\epsilon,U}$ from Eq. (2) we may simplify the right-hand side of the above inequality to arrive at

$$|f(U) - f(V)| \leq \frac{2\epsilon}{d} \sqrt{\sum_{z\in\mathcal{Z}} \frac{1}{w(z)} \left[\mathrm{Tr}\left(M_z (UQU^\dagger - VQV^\dagger)\right)\right]^2}.$$

It suffices to show that the sum in the square root is at most $O(d)\,\|U - V\|_{\mathrm{F}}^2$. Write $WDW^\dagger$ for the spectral decomposition of the Hermitian matrix $UQU^\dagger - VQV^\dagger$, where W is unitary and D is diagonal. As explained below, we have

$$\begin{aligned}
\sum_{z\in\mathcal{Z}} \frac{1}{w(z)} \left[\mathrm{Tr}\left(M_z W D W^\dagger\right)\right]^2 &= d^2 \sum_{z\in\mathcal{Z}} w(z) \left[\mathrm{Tr}\left(\left(\frac{W^\dagger M_z W}{w(z)d}\right) D\right)\right]^2 \\
&\leq d^2 \sum_{z\in\mathcal{Z}} w(z)\, \mathrm{Tr}\left(\left(\frac{W^\dagger M_z W}{w(z)d}\right) D^2\right) \\
&= d \sum_{z\in\mathcal{Z}} \mathrm{Tr}\left(W^\dagger M_z W D^2\right) \\
&= d \left\| UQU^\dagger - VQV^\dagger \right\|_{\mathrm{F}}^2 \\
&\leq 4d\, \|U - V\|_{\mathrm{F}}^2 \ ,
\end{aligned}$$

where in the second line we used the property that $WM_zW^\dagger/(w(z)d)$ is positive semidefinite with unit trace and applied Jensen's inequality to deduce that $(\mathrm{Tr}(AD))^2 = (\sum_i A_{ii} D_{ii})^2 \leq \sum_i A_{ii} D_{ii}^2 = \mathrm{Tr}(AD^2)$ for any positive semidefinite matrix A with unit trace. Also, in the fourth line we used the property that the measurement operators for the different outcomes z sum to identity. In the final line, we use the matrix inequality $\|AB\|_{\mathrm{F}} \leq \|A\|\, \|B\|_{\mathrm{F}}$ to deduce that

$$\begin{aligned}
\left\| UQU^\dagger - VQV^\dagger \right\|_{\mathrm{F}} &= \frac{1}{2} \left\| (U+V)Q(U-V)^\dagger + (U-V)Q(U+V)^\dagger \right\|_{\mathrm{F}} \\
&\leq \left\| (U+V)Q(U-V)^\dagger \right\|_{\mathrm{F}} \\
&\leq (\|UQ\| + \|VQ\|)\, \|U - V\|_{\mathrm{F}}
\end{aligned}$$

$$\leq 2 \|U - V\|_{\mathrm{F}} \ . \tag{26}$$

So $|f(U) - f(V)| \leq (4\epsilon/\sqrt{d}) \ \|U - V\|_{\mathrm{F}}$, i.e., f is $(4\epsilon/\sqrt{d})$-Lipschitz and Eq. (25) follows. The proof in the ℓ-outcome case is identical, except that the expectation is then

$$\mathbb{E}_{\boldsymbol{U} \sim \mathrm{Haar}} F^{\chi^2}_{\epsilon,d}(\mathcal{M}, \boldsymbol{U}) \leq \frac{4\ell\epsilon^2}{3d^2}$$

for $d \geq 2$, in accordance with the bound in Theorem 4.5. □

5.2 Sample complexity for adaptive measurements

Using the concentration of measure results derived in Section 5.1, we can show lower bounds for single-copy tomography robust to adaptively chosen measurements, so long as the number of different measurements that may be performed is suitably bounded. Our intermediate goal is to construct an ϵ-packing of states which are especially difficult to discriminate using the choice of measurements available to the learner. We invoke the tail bound from Lemma 5.3 to claim that for a non-negligible fraction of states of the form in Eq. (2), the measurement statistics from these measurements are uninformative. This is the content of the following lemma.

Lemma 5.4. *Fix an $\epsilon \in (0,1)$, positive integer $d \geq 4$, and a set of m measurements $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\} \subset \Xi(d)$. Let c, C be the universal constants defined in Lemma 5.3, and let $\alpha := c\epsilon^2/d$. For Haar-random $\boldsymbol{U} \in \mathsf{U}(d)$, the probability that $F^{\chi^2}_{\epsilon,d}(\mathcal{M}_i, \boldsymbol{U}) \leq \alpha + \epsilon^2 \ln(3m)/Cd^2$ for every $i \in [m]$ is at least $2/3$. Furthermore, if $\max_{i \in [m]} \mathrm{rank}(\mathcal{M}_i) = \ell$, the claim holds with $\alpha := 4\ell\epsilon^2/3d^2$.*

Proof. Applying the union bound over the m possible measurements we find that the probability that there is some measurement $i \in [m]$ such that $F^{\chi^2}_{\epsilon,d}(\mathcal{M}_i, \boldsymbol{U}) > \alpha + \epsilon^2 \ln(3m)/Cd^2$ is at most

$$\sum_{k=1}^{m} \Pr_{\boldsymbol{U} \sim \mathrm{Haar}} \left[F^{\chi^2}_{\epsilon,d}(\mathcal{M}_k, \boldsymbol{U}) > \alpha + \epsilon^2 \ln(3m)/Cd^2 \right] \leq m \exp\left(-\frac{Cd^2}{\epsilon^2} \cdot \frac{\epsilon^2 \ln(3m)}{Cd^2} \right) = \frac{1}{3} \ ,$$

where the inequality follows from the tail bound in Lemma 5.3. □

We now use a probabilistic existence argument to show that there is a packing which has the desired properties.

Corollary 5.5. *Fix an $\epsilon \in (0,1)$ and positive integer $d \geq 4$. Let $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\} \subset \Xi(d)$ be a fixed set of measurements and define α, C as in Lemma 5.4. There exists a set of N quantum states, $\mathcal{S} := \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ with*

$$\rho_i := \frac{2\epsilon}{d} U_i Q_{d/2} U_i^\dagger + (1 - \epsilon)\frac{\mathbb{1}}{d}$$

for some unitary operators $U_1, \ldots, U_N \in \mathsf{U}(d)$ such that

1. *$N \in \exp(\Omega(d^2))$,*
2. *$\mathcal{S}$ is an $(\epsilon/2)$-packing, and*
3. *$F^{\chi^2}_{\epsilon,d}(\mathcal{M}_i, U_j) \leq \alpha + \epsilon^2 \ln(3m)/Cd^2$ for every $i \in [m]$ and $j \in [N]$.*

Proof. The proof is similar to that of Corollary 3.4, except that it has an extra step. Suppose we have constructed a set of k quantum states $\mathcal{S}_k := \{\rho_1, \ldots, \rho_k\}$ with $k \leq \lfloor e^{d^2/32 - \ln(2)} \rfloor$, where the states are as in the statement of the corollary with corresponding unitary operators $\{U_1, \ldots, U_k\}$. Further suppose that $\mathcal{S}_k$ is an $\epsilon/2$-packing and that U_j satisfies the bound in part (3) of the statement for all $j \in [k]$. By setting the parameter $\xi := 1/2$ in Lemma 3.3 and making use of Lemma 5.4 and the union bound, we see that the probability of selecting a Haar-random unitary operator $\boldsymbol{U} \in \mathsf{U}(d)$ such that $\mathcal{S}_{k+1} := \mathcal{S}_k \cup \rho_{\epsilon,\boldsymbol{U}}$ no longer satisfies either condition is at most $1/2 + 1/3$. To be precise, the probability that $\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_j\right\|_1 \leq \epsilon/2$ for some $j \in [k]$ *or* that $F^{\chi^2}_{\epsilon,d}(\mathcal{M}_i, \boldsymbol{U}) > \alpha + \epsilon^2 \ln(3m)/Cd^2$ for some $i \in [m]$ is strictly less than one. Therefore, at least one state satisfying the desired properties exists, and the result follows by induction on k. □

We now have all the ingredients to derive the sample complexity for adaptive measurements.

Theorem 5.6. *Let $\epsilon \in (0,1)$. Any procedure for quantum tomography of d-dimensional quantum states that is $(\epsilon/2)$-accurate in trace distance and uses single-copy (possibly adaptive) measurements chosen from a fixed set of m measurements requires*

$$n \in \Omega\left(d^3 \left(1 + \log(m)/d\right)^{-1} / \epsilon^2\right)$$

samples of the unknown state.

Proof. Let $\mathcal{S} := \{\rho_1, \ldots, \rho_N\}$ be a set of $N \in \exp(\Omega(d^2))$ states which satisfies the conditions in Corollary 5.5 for the choice of m measurements $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\} \subset \Xi(d)$, with corresponding unitary operators $\{U_1, \ldots, U_N\} \subset \mathsf{U}(d)$. Let $\boldsymbol{x} \sim \mathrm{Unif}([N])$ and $\boldsymbol{y} := (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the measurement outcomes from applying n possibly adaptive measurements, each of which is an element of $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$, on identical copies of $\rho_{\boldsymbol{x}}$. (Recall that $\rho_{\boldsymbol{x}} = \rho_{\epsilon,U_{\boldsymbol{x}}}$.) By Fano's inequality as well as the assumption that the output of the tomography algorithm is accurate to within trace distance $\epsilon/2$, we have $I(\boldsymbol{x} : \boldsymbol{y}) \in \Omega(d^2)$.

On the other hand, we can upper bound the mutual information from above by using the properties of the states which comprise $\mathcal{S}$. Firstly, by the chain rule for mutual information we have

$$I(\boldsymbol{x} : \boldsymbol{y}) = \sum_{i=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i}) \tag{27}$$

where we use the shorthand $\boldsymbol{y}_{<i}$ to refer to the sequence of random variables $\boldsymbol{y}_{i-1}, \ldots, \boldsymbol{y}_1$. For each $i \in [n]$, let $p_{\boldsymbol{y}_i | y_{<i}, x}$ be the conditional distribution for the outcome of the i^{th} measurement performed on the i^{th} copy of the state ρ_x, given previous outcomes $y_{<i}$. The probabilities of this distribution are given by

$$p_{\boldsymbol{y}_i | y_{<i}, x}(y) := \mathrm{Tr}(M_y^{y_{<i}} \rho_x)$$

for each possible outcome y, where $\{M_y^{y_{<i}}\}_y$ is the POVM corresponding to the i^{th} measurement $\mathcal{M}^{y_{<i}}$ when the previous $i-1$ outcomes are $y_{<i}$. Also, let $w^{y_{<i}}(y) := \mathbb{E}_{\boldsymbol{U} \sim \mathrm{Haar}} \mathrm{Tr}(M_y^{y_{<i}} \rho_{\epsilon,\boldsymbol{U}})$ be a fixed distribution, for each possible sequence of prior outcomes $y_{<i}$. Consider the i^{th} term in the sum in the right-hand side of Eq. (27). We apply the upper bound on mutual information from Lemma 4.1 as well as Definition 5.2 for the function $F^{\chi^2}_{\epsilon,d}(\cdot,\cdot)$ to deduce that

$$I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i}) = \mathop{\mathbb{E}}_{\boldsymbol{y}'_{<i} \sim p_{\boldsymbol{y}_{<i}}} I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i} = \boldsymbol{y}'_{<i})$$

$$\leq \frac{1}{\ln(2)} \mathop{\mathbb{E}}_{\boldsymbol{y}'_{<i} \sim p_{\boldsymbol{y}_{<i}}} \mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_{\boldsymbol{x}|\boldsymbol{y}_{<i}}} \mathrm{D}_{\chi^2}(p_{\boldsymbol{y}_i|\boldsymbol{y}'_{<i},\boldsymbol{x}'} \,\|\, w^{\boldsymbol{y}'_{<i}})$$

$$= \frac{1}{\ln(2)} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<i}} \mathop{\mathbb{E}}_{\boldsymbol{x}|\boldsymbol{y}_{<i}} F^{\chi^2}_{\epsilon,d}(\mathcal{M}^{\boldsymbol{y}_{<i}}, U_{\boldsymbol{x}}) \tag{28}$$

$$\leq \frac{1}{\ln(2)} \left(c\epsilon^2/d + \epsilon^2 \ln(3m)/Cd^2\right)$$

$$\in O\left(\frac{\epsilon^2(1+\log(m)/d)}{d}\right). \tag{29}$$

where c, C are the universal constants defined in Lemma 5.4. The fourth line follows by the assumption that for every $y_{<i}$ we have $\mathcal{M}^{y_{<i}} = \mathcal{M}_j$ for some $j \in [m]$. Applying this argument to each of the n mutual information terms in Eq. (27) and combining with the relation $I(\boldsymbol{x} : \boldsymbol{y}) \in \Omega(d^2)$ gives the desired lower bound. □

For a fixed finite gate set, the number of distinct polylog(d)-size quantum circuits is at most polylog(d)$^{\text{polylog(d)}} \in \exp(o(d))$. Hence, a lower bound of $\Omega(d^3/\epsilon^2)$ samples holds in the setting where the learner is restricted to such circuits. This improves the bound we obtain from the Holevo theorem by a factor of d. Furthermore, this lower bound is tight by the algorithm we present in Appendix B.2, along with the fact that random Clifford circuits, which are efficiently implementable [AG04, VDB21], comprise a unitary 3-design [KG15, Web16, Zhu17]. The algorithm we present is nearly identical to an algorithm using Haar-random measurements given in Ref. [Wri16, Section 5.1], except we make use of the fact that measurements based on unitary 2-designs suffice[1].

Another particularly simple setting in which the above lower bound works well is that of d-outcome Pauli basis measurements, as considered by Yu [Yu20]. Yu shows a $\widetilde{O}(d^{3.32}/\epsilon^2)$ upper bound on the sample complexity of tomography with non-adaptive measurements, while Theorem 5.6 once again yields a lower bound of $\Omega(d^3/\epsilon^2)$, even with adaptive measurements.

We also have the following extension of the above theorem, which generalizes the lower bound for the case of two-outcome Pauli measurements due to Ref. [FGLE12] (although we do not consider possible dependence on the rank of the state here).

Theorem 5.7. *Let $\epsilon \in (0,1)$. Any procedure for quantum tomography of d-dimensional quantum states that is $(\epsilon/2)$-accurate in trace distance and uses single-copy (possibly adaptive), ℓ-outcome measurements chosen from a fixed set of m possible measurements requires*

$$n \in \Omega\left(d^4/(\ell + \log m)\epsilon^2\right)$$

samples of the unknown state.

Proof. The proof is identical to that for Theorem 5.6 except that the right-hand side of Eq. (28) is of the order of $4\ell\epsilon^2/3d^2 + \epsilon^2 \ln(3m)/Cd^2$, by Theorem 4.5. □

6 Sample complexity of classical shadows

In this section, we consider *classical shadows* and *shadow tomography*, variants of state tomography which have received much attention recently. Building on the ideas developed in the previous sections, we obtain new bounds on the sample complexity of these problems.

[1] Ref. [GKKT18] makes a similar observation; namely, that a measurement based on state 2-designs yields a sample complexity on the order of $d^3 \log(d)/\epsilon^2$.

6.1 Classical shadows

Full quantum state tomography is often unnecessary for determining important properties of a quantum system. For example, to verify the output of a quantum computer, one might only be concerned with comparing the state that is produced to some target pure state, perhaps by estimating their fidelity. Alternatively, in variational quantum algorithms an essential subroutine is to determine the expectation values of some observables encoding the cost function of interest. For both these tasks and more, succinctly represented information about the state known as a *classical shadow* [HKP20] can provide an exponential reduction in the number of copies of the state required to learn properties of interest. Informally, a classical shadow of a quantum state refers to a classical string, also called a *sketch*, using which we can estimate the expectation values of any given sequence of M observables to within accuracy ϵ. The sketch is produced by the measurement of individual copies of the otherwise unknown state.

For any state $\rho \in \mathsf{D}(d)$, consider the function f_ρ mapping $\mathsf{Psd}(d)^M$ to $\mathbb{R}^M$, defined as

$$f_\rho(E_1, E_2, \ldots, E_M) \coloneqq (\mathrm{Tr}(E_i\rho) : i \in [M]) \ .$$

More formally, the associated task is defined as defined below. In this definition, *single-copy access* refers to restricting measurements to individual copies of an unknown state, as described in Section 2.2.

Definition 6.1 (Classical shadows problem)**.** Given parameters $\epsilon \in (0,1)$, $B > 0$, and $M \geq 1$, and single-copy access to n copies of an unknown quantum state $\rho \in \mathsf{D}(d)$ the *classical shadows problem* consists of computing a description of a function $f : \mathsf{Psd}(d)^M \to \mathbb{R}^M$, called a *classical shadow*, such that for any fixed collection of M observables $(O_i : 0 \preceq O_i \preceq \mathbb{1},\ i \in [M])$ satisfying $\max_{i\in[M]} \mathrm{Tr}(O_i^2) = B$, it holds that $\left\|f(O_1,\ldots,O_M) - f_\rho(O_1,\ldots,O_M)\right\|_\infty \leq \epsilon$ with probability at least $2/3$.

Huang, Kueng, and Preskill [HKP20] give a procedure for computing classical shadows which uses only $n \in O(B\log(M)/\epsilon^2)$ efficient, nonadaptive measurements on single copies of the state ρ. Here, the measurements are implemented by using random q-qubit Clifford operators, which form a unitary 3-design. Then, the procedure performs a median-of-means estimation of the expectation values. Overall this is an unbounded improvement over full state tomography in the case where $\mathrm{Tr}(O_i^2)$ is at most a constant for the observables of interest O_i, since there is no explicit dependence on the dimension. They then show a matching lower bound in the nonadaptive measurement setting. However, this bound does not take into account the possibility of adaptive measurements. We turn to this in Section 6.2, focusing on the case where an upper bound on B is not known.

6.2 Lower bound with a limited choice of measurements

In this section, we show how the arguments developed in the previous sections for quantum tomography can be adjusted to give a lower bound for classical shadows with adaptive measurements, when the measurements are chosen from a "small enough" set. We obtain this result by proving the same lower bound for a variant of *shadow tomography* [Aar20] with single-copy measurements described below.

Definition 6.2 (Single-copy shadow tomography for bounded operators)**.** Given parameters $\epsilon \in (0,1)$ and $B > 0$, single-copy access to n copies of $\rho \in \mathsf{D}(2^q)$, as well as the description of M observables $(O_i : 0 \preceq O_i \preceq \mathbb{1},\ i \in [M])$ satisfying $\max_{i\in[M]} \mathrm{Tr}(O_i^2) = B$, the task is to output a vector $b \in \mathbb{R}^M$ such that with probability at least $2/3$ we have $|b_i - \mathrm{Tr}(O_i\rho)| \leq \epsilon$ for every $i \in [M]$.

	Upper bound	Lower bound
Entangled	$\widetilde{O}(\log(d)\log^2(M)/\epsilon^4)$ [BO21]	$\Omega(\log(M)/\epsilon^2)$ [Aar20]
Single-copy	$O(d\log(M)/\epsilon^2)$ [HKP20]	$\Omega(\min\{M/\log(M), d\}/\epsilon^2)$ [CCHL22]
Single-copy & Efficient	$O(d\log(M)/\epsilon^2)$ [HKP20]	$\Omega(d\log(M)/\epsilon^2)$ (this work)

Table 2: The best known upper and lower bounds on the sample complexity of shadow tomography for M observables $O_1, \ldots, O_M$, for $M \in \exp(O(d))$ for entangled measurements and $M \in \exp(O(d^2))$ for single-copy measurements. Note that for M larger than the corresponding thresholds, we may use state tomography with joint measurements or single-copy measurements to achieve sample complexity of order d^2/ϵ^2 and d^3/ϵ^2, respectively. The lower bounds for $M \in \exp(\Omega(d^2))$ are Ω of $d^2/\epsilon^2, d/\epsilon^2$, and d^3/ϵ^2, respectively for the three cases above. The $\widetilde{O}$ notation hides loglog factors in d and log factors in $1/\epsilon$.

Note that the output of the classical shadows problem can be used to produce a solution to the shadow tomography problem, when the Frobenius norm of the input operators is suitably bounded. Hence, any lower bound on the sample complexity for the latter task applies to the classical shadows problem as well.

Table 2 summarizes known results on the sample complexity of shadow tomography under various assumptions about the measurements. In Theorem 6.3 below, we prove a lower bound on the sample complexity of single-copy shadow tomography for bounded operators, when the possible measurements available to the learning algorithm are limited in number. The bound implies that in the setting of single-copy measurements with efficient circuits (i.e., uniformly generated polylog(d)-size quantum circuits over a finite universal gate set), the non-adaptive classical shadows algorithm due to Huang *et al.* [HKP20] is optimal for single-copy shadow tomography. In contrast with the lower bound due to Ref. [CCHL22] (in the second column of Table 2), a number of samples exponential in the number of qubits is inevitable using efficiently implementable measurements. This is a consequence of the fact that for a finite set of allowed measurements, one can always construct an instance of the classical shadows problem such that the measurement $\{O_i, \mathbb{1} - O_i\}$ is not in the set, for some $i \in [M]$.

Theorem 6.3. *Any algorithm for the classical shadows or the single-copy shadow tomography problem that only uses single-copy measurements chosen from a fixed set of m measurements requires*

$$\Omega\left(\frac{d\min\{d^2, \log M\}}{\epsilon^2(1+\log(m)/d)}\right)$$

samples when $B = d/2$.

Proof. As mentioned earlier, it suffices to prove the claimed lower bound for the shadow tomography problem. Consider any algorithm that only uses single-copy measurements chosen from a fixed set of m measurements $\{\mathcal{M}_1, \ldots, \mathcal{M}_m\} \subset \Xi(d)$. We construct a set of hard input instances for this algorithm and show that the algorithm requires a large number of samples for these instances.

We observe, as in Ref. [Aar20, Theorem 19], that well-separated states of the form we have been studying (cf. Eq. (2)) can be distinguished well by the measurements operators given by their deviation from the completely mixed state. We build on this to show that there exists a special collection of M states $\rho_1, \ldots, \rho_M$, and observables $O_1, \ldots, O_M$ whose expectation values enable us to

uniquely identify a state from the M alternatives $\rho_1, \ldots, \rho_M$. The states satisfy the additional property that the statistics obtained from measuring any of the ρ_i with any of m measurements $\mathcal{M}_j$ are not very informative. The lower bound then follows from Fano's inequality and the upper bound on the chi-squared divergence quantity we have been considering in the context of tomography. Since we may only take M to be at most $\exp(\kappa d^2)$ for a universal constant κ, the lower bound plateaus at this threshold.

More formally, we first construct the difficult instance of the shadow tomography problem. Let $\boldsymbol{U} \in \mathsf{U}(d)$ be a Haar-random unitary operator and, as before, let $Q \in \mathsf{Psd}(d)$ be a rank-$d/2$ orthogonal projection operator. By setting the parameter $t = 1/3$ in Lemma 3.2, we get that for any fixed rank-$d/2$ orthogonal projection operator $P \in \mathsf{Psd}(d)$,

$$\Pr[\mathrm{Tr}(P\boldsymbol{U}Q\boldsymbol{U}^\dagger) \geq d/3] \leq \exp(-c'd^2) \tag{30}$$

for a universal constant c'. Since the algorithm only uses single-copy measurements from a fixed set of m measurements, Lemma 5.4 applies and we have the following result.

Lemma 6.4. *Fix an $\epsilon \in (0,1)$ and a positive integer $d \geq 4$. Define α, C as in Lemma 5.4. There is a universal constant κ, such that for any $M \in [1, \exp(\kappa d^2)]$, there exists a set of M unitary operators $U_1, \ldots, U_M \in \mathsf{U}(d)$ such that*

1. $\mathrm{Tr}(U_i Q U_i^\dagger U_j Q U_j^\dagger) \leq d/3$ *for every $i, j \in [M]$, $i \neq j$, and*
2. $F^{\chi^2}_{\epsilon,d}(\mathcal{M}_i, U_j) \leq \alpha + \epsilon^2 \ln(3m)/Cd^2$ *for every $i \in [m]$ and $j \in [M]$.*

Proof. The proof is similar to that for Corollary 5.5 except that we use Eq. (30) to ensure the first condition (instead of using Lemma 5.4). □

Now, let $\mathcal{S} := \{\rho_1, \ldots, \rho_M\} \subset \mathsf{D}(d)$ be a collection of states of the form in Eq. (2), given by

$$\rho_i := \frac{2\epsilon}{d} U_i Q U_i^\dagger + (1-\epsilon)\frac{\mathbb{1}}{d} \ .$$

For any $i \in [M]$

$$\mathrm{Tr}(U_i Q U_i^\dagger \rho_i) = \frac{1}{2} + \frac{\epsilon}{2} \ ,$$

while by condition (1) in Lemma 6.4 we have for any $j \neq i$

$$\mathrm{Tr}(U_j Q U_j^\dagger \rho_i) \leq \frac{1}{2} + \frac{\epsilon}{6} \ .$$

This means that by estimating $\mathrm{Tr}(U_i Q U_i^\dagger \rho_x)$ with $\epsilon/12$ accuracy for every $i \in [M]$ we can identify the value of $x \in [M]$. Thus, we may use the algorithm for shadow tomography with input observables $U_1 Q U_1^\dagger, \ldots, U_M Q U_M^\dagger$ to discriminate between the M states in $\mathcal{S}$ with probability at least $2/3$. We argue next that the algorithm requires a "large" number of single-copy measurements in order to accomplish this.

Let $\boldsymbol{x}$ be uniformly random over $[M]$ and $\boldsymbol{y} := (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the measurement outcomes obtained from n single-copy (possibly adaptive) measurements performed on distinct copies of the state $\rho_{\boldsymbol{x}}$. By chain rule for mutual information, we have

$$I(\boldsymbol{x} : \boldsymbol{y}) = \sum_{i=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i})$$

$$\leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<i}} \mathop{\mathbb{E}}_{\boldsymbol{x}|\boldsymbol{y}_{<i}} F_{\epsilon,d}^{\chi^2}(\mathcal{M}^{\boldsymbol{y}_{<i}}, U_{\boldsymbol{x}})$$
$$\in O\left(\frac{n\epsilon^2(1+\log(m)/d)}{d}\right) , \tag{31}$$

where we have omitted some steps since they are identical to those leading to Eq. (29).

On the other hand, since the algorithm identifies the state $\rho_{\boldsymbol{x}}$ with probability $\geq 2/3$ from the measurement outcomes $\boldsymbol{y}$, by Fano's Inequality, $I(\boldsymbol{x}:\boldsymbol{y}) \geq \Omega(\log(M))$. This concludes the proof of Theorem 6.3. □

Suppose an algorithm for classical shadows or shadow tomography uses only efficient single-copy measurements over a fixed, finite universal gate set. The number m of different measurements it may use is then $O(\exp(\text{polylog}(d)))$. (See Appendix C for further justification and a slight generalization.) By the above lower bound, the algorithm requires $\Omega(d\min\{d^2, \log(M)\}/\epsilon^2)$ samples. In fact, this bound is optimal.

Lemma 6.5. *There is an algorithm that uses only efficient, single-copy measurements and*

$$O(d\min\{d^2, \log(M)\}/\epsilon^2)$$

samples and solves the classical shadows and single-copy shadow tomography problems for arbitrary B.

Proof. The upper bound is achieved by the following procedure: use the random Clifford operator-based classical shadows algorithm from Ref. [HKP20, Theorem 1] if $M \leq e^{d^2}$, and the random Clifford operator-based state tomography algorithm of Ref. [KRT17] otherwise. The former has sample complexity of the order of $d\log(M)/\epsilon^2$, and the latter d^3/ϵ^2. □

6.3 Sample means suffice for single-copy shadow tomography

We turn our attention to the case of nonadaptive — but otherwise arbitrary — single-copy measurements on an unknown state $\rho \in \mathsf{D}(d)$. As can be seen from Table 2, the median-of-means algorithm due to Ref. [HKP20] is optimal up to log factors for shadow tomography in this setting. Their proposal employs random Clifford operations to perform random basis measurements. However, we may take an even simpler approach using the same measurement scheme, which also turns out to be optimal. Specifically, we show that taking the sample means using the classical shadow reproduces the same upper bound on the overall sample complexity, which is $n \in O(\min\{d, M\}\log(M)/\epsilon^2)$, assuming $M \leq e^{d^2}$.

We first handle the case when $M > d$. Suppose we apply a random Clifford operator $\boldsymbol{U} \in \mathsf{U}(d)$ and then measure in the standard basis $\{|j\rangle\}_{j=1}^{d} \subset \mathbb{C}^d$. For a fixed Clifford operator $U \in \mathsf{U}(d)$ we may write the operators for this measurement as $\{U|j\rangle\langle j|U^\dagger\}_{j=1}^{d}$. It is well-known that this random projective measurement is closely related to state t-designs, as explained in Appendix 6.3. Define the random variable $\hat{\rho}(\boldsymbol{U}, \boldsymbol{j}) = (d+1)\boldsymbol{U}|\boldsymbol{j}\rangle\langle\boldsymbol{j}|\boldsymbol{U}^\dagger - \mathbb{1}$ where $\boldsymbol{j} \in [d]$ is the random measurement outcome in the standard basis. By Proposition B.1 in Appendix B we have $\mathbb{E}\,\hat{\rho}(\boldsymbol{U}, \boldsymbol{j}) = \rho$. We also make use of the following property.

Proposition 6.6 (Prop. S1, Sec. 5 in the supplementary materials for Ref. [HKP20])**.** *Let X be a Hermitian operator with $-\mathbb{1} \preceq X \preceq \mathbb{1}$ acting on $\mathbb{C}^d$, and let $\hat{\rho}(\boldsymbol{U}, \boldsymbol{j})$ be as defined above. It holds that*

$$\text{Var}\left[\text{Tr}(X\hat{\rho}(\boldsymbol{U}, \boldsymbol{j}))\right] \leq 3\,\text{Tr}(X^2).$$

Finally, we require a concentration of measure property of bounded random variables known as Bernstein's inequality. This is stronger than Hoeffding's inequality when the variances of the random variables are sufficiently small. This version of Bernstein's inequality can be found in Ref. [Ver18], for example.

Theorem 6.7 (Theorem 2.8.4 in Ref. [Ver18]). *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent, mean zero random variables such that $|\boldsymbol{x}_i| \leq K$ with probability 1 for all $i \in [n]$. Then, for every $\epsilon \geq 0$, we have*

$$\Pr\left[\left|\sum_{i=1}^{n} \boldsymbol{x}_i\right| \geq \epsilon\right] \leq 2\exp\left(\frac{-\epsilon^2/2}{\sigma^2 + K\epsilon/3}\right)$$

where $\sigma^2 := \sum_{i=1}^{n} \mathbb{E}\, \boldsymbol{x}_i^2$.

Now suppose that the observables given as input to the shadow tomography algorithm are O_i, with $0 \preceq O_1, \ldots, O_M \preceq \mathbb{1}$. Define the random variables $f_i(\boldsymbol{U}, \boldsymbol{j}) := \mathrm{Tr}(O_i \hat{\rho}(\boldsymbol{U}, \boldsymbol{j}))$ for each $i \in [M]$. It holds that

$$\mathbb{E}\, f_i(\boldsymbol{U}, \boldsymbol{j}) = \mathrm{Tr}(O_i\, \mathbb{E}\, \hat{\rho}(\boldsymbol{U}, \boldsymbol{j})) = \mathrm{Tr}(O_i \rho) \ , \tag{32}$$

so that $f_i(\boldsymbol{U}, \boldsymbol{j})$ is an unbiased estimator for $\mathrm{Tr}(O_i\rho)$. If we perform the random measurement described above on n separate copies of ρ, we obtain i.i.d. random variables $(\boldsymbol{U}_1, \boldsymbol{j}_1), \ldots, (\boldsymbol{U}_n, \boldsymbol{j}_n)$. These define the classical shadow of the state as

$$\frac{1}{n} \sum_{k=1}^{n} \hat{\rho}(\boldsymbol{U}_k, \boldsymbol{j}_k) \ .$$

The expectation value for O_i predicted by the classical shadow is the sample mean of the i^{th} estimator f_i. For any $\epsilon > 0$, by Bernstein's inequality we have that

$$\Pr\left[\left|\frac{1}{n}\sum_{k=1}^{n} f_i(\boldsymbol{U}_k, \boldsymbol{j}_k) - \mathrm{Tr}(O_i\rho)\right| > \epsilon\right] \leq 2\exp\left(\frac{-\epsilon^2/2}{\sigma^2 + \epsilon K/(3n)}\right)$$

where $\sigma^2 := \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}[f_i(\boldsymbol{U}_k, \boldsymbol{j}_k)]$ and K is such that $|f_i(\boldsymbol{U}_k, \boldsymbol{j}_k) - \mathrm{Tr}(O_i\rho)| \leq K$ with probability 1 for all $k \in [n]$. By definition $\|f_i\|_\infty \leq d + 1$ so K can be taken to be $O(d)$, and by Proposition 6.6 we have $\sigma^2 \leq 3d/n$. Taking $n \in O(d \log(M)/\epsilon^2)$, the probability above is at most $1/3M$. By the union bound, we may estimate $\mathrm{Tr}(O_i\rho)$ for all $i \in [M]$ to additive error ϵ using these n samples, with failure probability at most $1/3$. We remark that in the setting where the measurements used by the algorithm are efficient, this describes the optimal procedure.

Consider $M \leq d$. In this case, we perform the two-outcome measurement with operators $\{O_i, \mathbb{1} - O_i\}$ a total of $O(\log(M)/\epsilon^2)$ times for each $i \in [M]$. The sample means are then within ϵ of the corresponding expectation values. The procedure uses $O(M \log(M)/\epsilon^2)$ samples of the state, and matches the information-theoretic lower bound proved in Ref. [CCHL22] up to a factor of $\log^2(M)$ (see the second column of Table 2).

7 Open problems

We conclude with some directions for future work arising from the lower bounds in Sections 5 and 6. In Theorem 5.7 we incur a polylog(d) factor in the denominator of the lower bound

for tomography with efficient, constant-outcome, single-copy measurements. Can this be improved? Note that such a factor also appears in the denominator of the lower bound for binary Pauli measurements in Ref. [FGLE12]. Is there a way to incorporate rank-dependence into the lower bounds appearing in Section 5 for adaptive tomography with limited measurement settings? The approach we took to incorporate the dependence on the norm parameter B into the lower bounds for classical shadows does not carry over well to the setting of rank-dependent quantum tomography, since the packing we constructed has states with rank up to d. Finally, are there simpler information-theoretic arguments that yield the unconditional bounds obtained in Refs. [CHLL22, CHL+22]?

References

[Aar20] Scott Aaronson. Shadow tomography of quantum states. *SIAM Journal on Computing*, 49(5):STOC18–368–STOC18–394, 2020.

[ACH+19] Scott Aaronson, Xinyi Chen, Elad Hazan, Satyen Kale, and Ashwin Nayak. Online learning of quantum states. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124019, December 2019.

[AE07] Andris Ambainis and Joseph Emerson. Quantum t-designs: t-wise independence in the quantum world. In *2007 22nd Annual IEEE Conference on Computational Complexity*, pages 129–140, Los Alamitos, CA, USA, June 2007. IEEE Computer Society.

[AG04] Scott Aaronson and Daniel Gottesman. Improved simulation of stabilizer circuits. *Physical Review A*, 70(5), November 2004.

[ALL22] Anurag Anshu, Zeph Landau, and Yunchao Liu. Distributed quantum inner product estimation. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, pages 44–51, New York, NY, USA, 2022. Association for Computing Machinery.

[AW02] Rudolph Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[BCL20] Sebastien Bubeck, Sitan Chen, and Jerry Li. Entanglement is necessary for optimal quantum property testing. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 692–703, Los Alamitos, CA, USA, November 2020. IEEE Computer Society.

[BD10] Francesco Buscemi and Nilanjana Datta. The quantum capacity of channels with arbitrarily correlated noise. *IEEE Transactions on Information Theory*, 56(3):1447–1460, 2010.

[BO21] Costin Bădescu and Ryan O'Donnell. Improved quantum data analysis. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pages 1398–1411, New York, NY, USA, 2021. Association for Computing Machinery.

[CCHL21] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. A hierarchy for replica quantum advantage. Technical Report arXiv:2111.05874 [quant-ph], arXiv, http://www.arxiv.org/, 2021.

[CCHL22] Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 574–585, Los Alamitos, CA, USA, February 2022. IEEE Computer Society.

[CD10] Jing Chen and Andrew Drucker. Short multi-prover quantum proofs for SAT without entangled measurements. Technical Report arXiv:1011.0716 [quant-ph], arXiv, http://www.arxiv.org/, 2010.

[CHL+22] Sitan Chen, Brice Huang, Jerry Li, Allen Liu, and Mark Sellke. Tight bounds for state tomography with incoherent measurements. Technical Report arXiv:2206.05265 [quant-ph], arXiv, http://www.arxiv.org/, 2022.

[CHLL22] Sitan Chen, Brice Huang, Jerry Li, and Allen Liu. Tight bounds for quantum state certification with incoherent measurements. Technical Report arXiv:2204.07155 [quant-ph], arXiv, 2022.

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, April 2005.

[Fan66] Robert M. Fano. *Transmission of Information: a Statistical Theory of Communications*. MIT Press, 1966.

[FGLE12] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, September 2012.

[GKKT18] Madalin Guta, Jonas Kahn, Richard Kueng, and Joel A. Tropp. Fast state tomography with optimal error bounds, 2018.

[GKKT20] Madalin Guţă, Jonas Kahn, Richard Kueng, and Joel A. Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, April 2020.

[HH12] Ferenc Huszár and Neil M. T. Houlsby. Adaptive bayesian quantum tomography. *Physical Review A*, 85:052120, May 2012.

[HHJ+17] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63:5628–5641, September 2017.

[HKP20] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, June 2020.

[HKP21] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Physical Review Letters*, 126:190505, May 2021.

[HLW06] Patrick Hayden, Debbie W. Leung, and Andreas Winter. Aspects of generic entanglement. *Communications in Mathematical Physics*, 265(1):95–117, March 2006.

[HMR+10] Sean Hallgren, Cristopher Moore, Martin Rötteler, Alexander Russell, and Pranab Sen. Limitations of quantum coset states for graph isomorphism. *Journal of the ACM*, 57(6):1–33, November 2010.

[HRS05] Sean Hallgren, Martin Roetteler, and Pranab Sen. Limitations of quantum coset states for graph isomorphism. Technical Report arXiv:quant-ph/0511148, arXiv, http://www.arxiv.org/, 2005.

[KG15] Richard Kueng and David Gross. Qubit stabilizer states are complex projective 3-designs. Technical Report arXiv:1510.02767 [quant-ph], arXiv, http://www.arxiv.org/, 2015.

[KRT17] 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.

[LN22] Angus Lowe and Ashwin Nayak. Improved lower bounds for learning quantum states with unentangled measurements. Presented at the 25th Annual Conference on Quantum Information Processing, March 7–12, 2022.

[Low21] Angus Lowe. Learning quantum states without entangled measurements. M.Math. thesis, University of Waterloo, Waterloo, Ontario, Canada, October 2021.

[Mec19] Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, July 2019.

[MRD+13] D. H. Mahler, Lee A. Rozema, Ardavan Darabi, Christopher Ferrie, Robin Blume-Kohout, and Aephraim M. Steinberg. Adaptive quantum state tomography improves accuracy quadratically. *Physical Review Letters*, 111:183601, October 2013.

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 10th anniversary edition edition, 2010.

[OW16] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 899–912, New York, NY, USA, 2016. Association for Computing Machinery.

[OW17] 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, pages 962–974, New York, NY, USA, 2017. Association for Computing Machinery.

[SV16] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[SZK+21] G.I. Struchalin, Ya. A. Zagorovskii, E.V. Kovlakov, S.S. Straupe, and S.P. Kulik. Experimental estimation of quantum state properties from classical shadows. *PRX Quantum*, 2:010307, January 2021.

[Tsy09] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, New York, NY, 2009.

[VDB21] Ewout Van Den Berg. A simple method for sampling random clifford operators. In *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pages 54–59, 2021.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[Wat18] John Watrous. *The Theory of Quantum Information*. Cambridge University Press, 2018.

[Web16] Zak Webb. The clifford group forms a unitary 3-design. *Quantum Information & Computation*, 16(15-16):1379–1400, November 2016.

[Wri16] John Wright. *How to Learn a Quantum State*. PhD thesis, Carnegie Mellon University, 2016.

[Yu20] Nengkun Yu. Sample efficient tomography via Pauli measurements. Technical Report arXiv:2009.04610 [quant-ph], arXiv, http://www.arxiv.org/, 2020.

[Zhu17] Huangjun Zhu. Multiqubit clifford groups are unitary 3-designs. *Physical Review A*, 96:062336, December 2017.

A Haar integrals

The Haar measure μ is the unique unitarily invariant probability measure on the space of unitary operators, $\mathsf{U}(d)$. Using this measure, one may define channels $\Phi_k : (\mathbb{C}^{d\times d})^{\otimes k} \to (\mathbb{C}^{d\times d})^{\otimes k}$ of the form

$$\Phi_k(X) = \int_{\mathsf{U}(d)} U^{\otimes k} X (U^\dagger)^{\otimes k} \mathrm{d}\mu(U), \tag{33}$$

which are referred to as "twirl" operations. In the rest of this section, we evaluate this channel explicitly in the case where the operator X is a tensor product of orthogonal projectors onto subspaces of $\mathbb{C}^d$. Following the presentation in Ref. [Wat18], we make use of an important result on the structure of permutation-invariant operators. Recall from Section 2 that S_k is the symmetric group on $\{1,\ldots,k\}$ and W_π is the operator on $(\mathbb{C}^{d\times d})^{\otimes k}$ that permutes the k tensor factors according to the permutation $\pi \in S_k$.

Theorem A.1 (Theorem 7.15 in Ref. [Wat18])**.** *Let $k > 0$ be a positive integer and $X \in (\mathbb{C}^{d\times d})^{\otimes k}$ be an operator. The following are equivalent:*

1. $[X, U^{\otimes k}] = 0 \ \forall U \in \mathsf{U}(d)$.
2. $X = \sum_{\pi\in S_k} v(\pi) W_\pi$ *for some choice of* $v \in \mathbb{C}^{|S_k|}$.

Since $\Phi_k(X)$ satisfies the first condition, we can apply the theorem to write the output of the channel as a linear combination of permutation operators. This helps us evaluate the Haar integrals which arise in this work.

Proposition A.2. *Let $d > 1$ be a positive integer, $Q \in \mathsf{Psd}(d)$ a rank-r orthogonal projection operator, and $\boldsymbol{U} \in \mathsf{U}(d)$ a Haar-random unitary operator. It holds that*

$$\mathbb{E}\, \boldsymbol{U} Q \boldsymbol{U}^\dagger = \frac{r\mathbb{1}}{d}.$$

Proof. We can write the expectation as

$$\int_{\mathsf{U}(d)} UQU^\dagger \mathrm{d}\mu(U) = \Phi_1(Q).$$

By Theorem A.1 we have

$$\mathbb{E}\, \boldsymbol{U} Q \boldsymbol{U}^\dagger = \kappa \mathbb{1}$$

where $\kappa \in \mathbb{C}$ is some coefficient depending on Q. Recalling that Q is a rank-r orthogonal projection operator, taking the trace of both sides and solving for κ yields $\kappa = r/d$. □

Proposition A.3. *Let $d > 1$ be a positive integer. Let $\Pi_1, \Pi_2 \in \mathsf{Psd}(d)$ be orthogonal projection operators of rank r_1, r_2, respectively, such that the image of Π_1 is contained in that of Π_2. For $\boldsymbol{U} \in \mathsf{U}(d)$ a Haar-random unitary operator it holds that*

$$\mathbb{E}\, \boldsymbol{U}^{\otimes 2}(\Pi_1 \otimes \Pi_2)(\boldsymbol{U}^\dagger)^{\otimes 2} = \frac{r_1}{d(d^2-1)}\left[(r_2 d - 1)\mathbb{1} + (d - r_2)W\right]$$

where W is the swap operator acting on $(\mathbb{C}^d)^{\otimes 2}$.

Proof. We can write the expectation as

$$\int_{\mathsf{U}(d)} U^{\otimes 2}(\Pi_1 \otimes \Pi_2)(U^\dagger)^{\otimes 2} \mathrm{d}\mu(U) = \Phi_2(\Pi_1 \otimes \Pi_2).$$

By Theorem A.1 we have

$$\mathbb{E}\, \boldsymbol{U}^{\otimes 2}(\Pi_1 \otimes \Pi_2)(\boldsymbol{U}^\dagger)^{\otimes 2} = \alpha \mathbb{1} \otimes \mathbb{1} + \beta W$$

where W is the swap operator and $\alpha, \beta \in \mathbb{C}$ are some coefficients depending on Q. Left-multiplying by $\mathbb{1} \otimes \mathbb{1}$ or W and taking the trace of both sides yields

$$\mathrm{Tr}(\Pi_1 \otimes \Pi_2) = r_1 r_2 = \alpha d^2 + \beta d, \quad \mathrm{Tr}(W(\Pi_1 \otimes \Pi_2)) = r_1 = \alpha d + \beta d^2 \ ,$$

as $\Pi_1 \Pi_2 = \Pi_1$. This allows us to solve for α, β:

$$\alpha = \frac{r_1(r_2 d - 1)}{d(d^2-1)}, \quad \beta = \frac{r_1(d - r_2)}{d(d^2-1)}. \tag{34}$$

This concludes the proof of the proposition. □

We also make use of the expectations of operators of the following form.

Proposition A.4. *Let $d \geq 1$ and $\boldsymbol{U}$ be a Haar-random unitary operator over $\mathbb{C}^d$. For any $i, j \in [d]$, we have*

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{U}} \boldsymbol{U}|i\rangle \otimes \boldsymbol{U}|j\rangle &= 0 \ , \\
\mathop{\mathbb{E}}_{\boldsymbol{U}} \langle i|\boldsymbol{U}^\dagger \otimes \langle j|\boldsymbol{U}^\dagger &= 0 \ , \\
\mathop{\mathbb{E}}_{\boldsymbol{U}} \boldsymbol{U}|i\rangle \otimes \langle j|\boldsymbol{U}^\dagger &= \frac{\delta_{ij}}{d} \sum_{k=1}^{d} |k\rangle \otimes \langle k| \ , \qquad \textit{and} \\
\mathop{\mathbb{E}}_{\boldsymbol{U}} \langle j|\boldsymbol{U}^\dagger \otimes \boldsymbol{U}|i\rangle &= \frac{\delta_{ij}}{d} \sum_{k=1}^{d} \langle k| \otimes |k\rangle \ ,
\end{aligned}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

Proof. The second identity follows from the first by taking the adjoint, which commutes with taking the expectation over $\boldsymbol{U}$. Similarly, the fourth identity follows from the third by conjugating with the swap operator on $\mathbb{C}^d \otimes \mathbb{C}^d$.

The first identity follows from the invariance of the Haar measure under multiplication by $\mathrm{i}\mathbb{1}$; we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \boldsymbol{U}|i\rangle \otimes \boldsymbol{U}|j\rangle = \mathrm{i}^2 \mathop{\mathbb{E}}_{\boldsymbol{U}} \boldsymbol{U}|i\rangle \otimes \boldsymbol{U}|j\rangle = 0 \ .$$

Similarly, if $i \neq j$, then by the invariance of the Haar-measure under multiplication on the right by the unitary operator $\mathbb{1} - 2|j\rangle\langle j|$, the third identity holds.

Let A be the left hand side of the third identity when $i = j$. Then $\langle k|A|l\rangle = 0$ if $k \neq l$, by the invariance of the Haar-measure under multiplication on the right by the operator $\mathbb{1} - 2|l\rangle\langle l|$. Furthermore, $\langle k|A|k\rangle = \mathbb{E}_{\boldsymbol{U}} |\langle k|\boldsymbol{U}|i\rangle|^2 = 1/d$, by the invariance of the Haar measure under permutations of the standard basis elements. □

B Algorithms for quantum tomography

B.1 Tomography with entangled measurements

In the entangled measurement model, it has been shown by O'Donnell and Wright [OW16] and Haah et al. [HHJ+17] that $O(d^2/\epsilon^2)$ copies of the state suffice to estimate it to ϵ-accuracy in trace distance with high probability[2]. At the same time, a matching lower bound was also shown in [HHJ+17]. So the sample complexity of tomography in the entangled measurement setting is known up to a constant factor, for constant probability of success. A full description of these algorithms is outside the scope of this work, requiring ideas from representation theory and in particular the relationship between certain representations on $(\mathbb{C}^d)^{\otimes n}$. We refer the interested reader to Chapters 2 and 5 of the Wright's PhD thesis [Wri16].

B.2 Tomography with random basis measurements

For completeness we describe an algorithm which achieves a sample complexity of $O(d^3/\epsilon^2)$ for ϵ-accurate tomography (in trace distance) using efficiently implementable, nonadaptive measurements. The analysis we present is due to Wright [Wri16, Section 5.1], with minor differences. We also point out that measurement based on a state 2-design suffices. These may be derived from a *spherical* 4-*design* or a unitary 2-design.

An algorithm for the bounded-rank case follows from Ref. [KRT17, Theorem 2]. Haah *et al.* sketch the details of this algorithm in Ref. [HHJ+17, Section II.A]. They invoke an "operator Chernoff bound" due to Ahlswede and Winter [AW02] to conclude that the sample average of m i.i.d. standard normal vectors $|\boldsymbol{\psi}_i\rangle \in \mathbb{C}^d$ with $|\boldsymbol{\psi}_i\rangle \sim \mathrm{N}(0, \mathbb{1})$ approximates the identity operator $\mathbb{1}$. Formally, we have

$$\left\| \frac{1}{m} \sum_{i=1}^{m} |\boldsymbol{\psi}_i\rangle\langle\boldsymbol{\psi}_i| - \mathbb{1} \right\| \leq \alpha \ , \tag{35}$$

for a constant $\alpha > 0$, with probability at least $3/4$, provided $m \in \Omega(d(\ln d)/\alpha^2)$. This leads to a sample complexity of $O(r^2 d/\epsilon^2)$ for $r \geq \ln d$, and $O(rd(\ln d)/\epsilon^2)$ for $r \leq \ln d$. A stronger tail inequality [Ver18, Theorem 4.6.1] guarantees that Eq. (35) holds for a suitable constant α, with

[2]Originally, the upper bound presented in Haah et al. [HHJ+17] had an additional factor of $\log(d/\epsilon)$, which was subsequently removed in the thesis of Wright [Wri16].

probability at least $1 - 2\exp(-m)$ as long as $m \geq d$. This gives us the optimal bound of $O(r^2 d/\epsilon^2)$ on the sample complexity of the algorithm. Guţă, Kahn, Kueng, and Tropp [GKKT20, Theorem 2] give a different algorithm that also achieves the optimal sample complexity.

Let $\rho \in \mathsf{D}(d)$ be the state to be learned, and $\{|j\rangle\}_{j=1}^d$ be the standard basis. Consider sampling a random unitary operator $\boldsymbol{U}$ comprising a unitary 2-design and then performing the basis measurement corresponding to the measurement operators $\{\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger\}_{j=1}^d$, obtaining outcome $\boldsymbol{j}$. Suppose we do this on n separate copies of the state, resulting in iid random variables $(\boldsymbol{U}_1, \boldsymbol{j}_1), \ldots, (\boldsymbol{U}_n, \boldsymbol{j}_n)$ where $\boldsymbol{U}_i$ is the i^{th} random unitary operator and $\boldsymbol{j}_i$ is the outcome from the i^{th} measurement. Define $\hat{\rho}(U, j) := (d+1)U|j\rangle\langle j|U^\dagger - \mathbb{1}$ for $U \in \mathsf{U}(d)$ and $j \in [d]$.

Proposition B.1. *It holds that*

$$\mathbb{E}\,\hat{\rho}(\boldsymbol{U}, \boldsymbol{j}) = \rho.$$

Proof. Let $p_{\boldsymbol{U}}$ denote the distribution of $\boldsymbol{U}$ and $p_{j|U}(j)$ the probability of obtaining outcome j given that U is drawn. We have

$$\begin{aligned}\mathbb{E}\,\boldsymbol{U}|\boldsymbol{j}\rangle\langle\boldsymbol{j}|\boldsymbol{U}^\dagger &= \sum_{j=1}^d \mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} p_{j|\boldsymbol{U}}(j)\; \boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger \\ &= \sum_{j=1}^d \mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} \langle j|\boldsymbol{U}\rho\boldsymbol{U}^\dagger|j\rangle\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger. \end{aligned} \tag{36}$$

Consider the jth term in the sum above. We may write that term equivalently as

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} \mathrm{Tr}_2\left((\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger)^{\otimes 2}(\mathbb{1}\otimes\rho)\right) = \mathrm{Tr}_2\left(\mathop{\mathbb{E}}_{V\sim\mathrm{Haar}}(\boldsymbol{V}|j\rangle\langle j|\boldsymbol{V}^\dagger)^{\otimes 2}(\mathbb{1}\otimes\rho)\right) \tag{37}$$

where the equality follows from linearity of trace and the choice of $\boldsymbol{U}$ as a 2-design. Note that it suffices that the measurement operators be derived from a state 2-design (see, e.g., Ref. [AE07]). Proposition A.3 gives an explicit solution to the Haar integral inside the partial trace for the general case of a rank-r projector rather than $|j\rangle\langle j|$. Taking $r = 1$, we find that

$$\mathop{\mathbb{E}}_{V\sim\mathrm{Haar}}(\boldsymbol{V}|j\rangle\langle j|\boldsymbol{V}^\dagger)^{\otimes 2} = \frac{1}{d(d+1)}\left[\mathbb{1}\otimes\mathbb{1} + W\right].$$

Substituting into the right-hand side of Eq. (37) and making use of the identities $\mathrm{Tr}_2(W(\mathbb{1}\otimes\rho)) = \rho$ and $\mathrm{Tr}(\rho) = 1$ we find that it is equal to $\frac{1}{d(d+1)}(\mathbb{1}+\rho)$. Using the property that this holds for any $j \in [d]$ and substituting into Eq. (36) we obtain the relation $\mathbb{E}\,\boldsymbol{V}|\boldsymbol{j}\rangle\langle\boldsymbol{j}|\boldsymbol{V}^\dagger = \frac{1}{d+1}(\mathbb{1}+\rho)$. The proposition then follows from the definition of $\hat{\rho}(\boldsymbol{U}, \boldsymbol{j})$. □

In other words, $\hat{\rho}(\boldsymbol{U}, \boldsymbol{j})$ is an unbiased estimator of ρ. We take the empirical average of the n independent samples of this estimator $\frac{1}{n}\sum_{i=1}^n \hat{\rho}(\boldsymbol{U}_i, \boldsymbol{j}_i)$ which we obtained by measuring n separate copies of the state. Then the squared distance between the estimator and the true state in terms of the metric induced by the Frobenius norm is

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \hat{\rho}(\boldsymbol{U}_i, \boldsymbol{j}_i) - \rho\right\|_{\mathrm{F}}^2 &= \frac{1}{n^2}\,\mathbb{E}\left\|\sum_{i=1}^n (\hat{\rho}(\boldsymbol{U}_i, \boldsymbol{j}_i) - \rho)\right\|_{\mathrm{F}}^2 \\ &= \frac{1}{n^2}\,\mathrm{Tr}\left(\mathbb{E}\left[\sum_{i=1}^n (\hat{\rho}(\boldsymbol{U}_i, \boldsymbol{j}_i) - \rho)\right]^2\right).\end{aligned}$$

It is straightforward to show that for a sum of n mean-zero, independent random matrices $\boldsymbol{A}_i$ it holds that $\mathbb{E}\left[\sum_{i=1}^n \boldsymbol{A}_i\right]^2 = \sum_{i=1}^n \mathbb{E}\,\boldsymbol{A}_i^2$, which entails that the right-hand side of the above is

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^n \mathrm{Tr}\left(\mathbb{E}(\hat{\rho}(\boldsymbol{U}_i,\boldsymbol{j}_i)-\rho)^2\right) &= \frac{1}{n^2}\sum_{i=1}^n \left(\mathbb{E}\,\mathrm{Tr}(\hat{\rho}(\boldsymbol{U}_i,\boldsymbol{j}_i)^2) - \mathrm{Tr}(\rho^2)\right)\\ &\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\,\mathrm{Tr}(\hat{\rho}(\boldsymbol{U}_i,\boldsymbol{j}_i)^2)\\ &= \frac{d^2+d-1}{n}\end{aligned}$$

where the inequality used $\mathrm{Tr}(\rho^2) \geq 0$ and the final line comes from the following calculation. For a Hermitian matrix A, we have $\mathrm{Tr}(A^2) = \sum_{i=1}^d \lambda_i(A)^2$. In our case, all eigenvalues of the operator $(d+1)U|j\rangle\langle j|U^\dagger - \mathbb{1}$ except one are -1, and one eigenvalue is d. Using the matrix inequality $\|\cdot\|_1 \leq \sqrt{d}\,\|\cdot\|_{\mathrm{F}}$, we obtain the inequality

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \hat{\rho}(\boldsymbol{U}_i,\boldsymbol{j}_i) - \rho\right\|_1^2 \leq \frac{d(d^2+d-1)}{n}.$$

Substituting $n \in O(d^3/\epsilon^2)$ gives us the desired upper bound on error in expectation. We can achieve error at most ϵ with high (constant) probability using Markov's inequality, with a constant factor increase in the number of samples.

B.3 Tomography with binary Pauli measurements

In the setting of binary Pauli measurements there exists perhaps the most straightforward tomography algorithm, to the point where its $O(d^4/\epsilon^2)$ sample complexity is folklore. However, since we show that this is the information-theoretically optimal algorithm for a class of nonadaptive measurement scenarios, it may be worth reviewing. The general q-qubit Pauli matrices are the various Hermitian, unitary, and traceless q-fold tensor products of the set of single-qubit Pauli matrices $\{\mathbb{1},\sigma_x,\sigma_y,\sigma_z\} \subset \mathbb{C}^{2\times 2}$. This means that there are $4^q = d^2$ different q-qubit Pauli matrices $\mathcal{P}_d = \{P_1,\ldots,P_{d^2}\}$, where we let $d = 2^q$. These operators form an orthogonal basis for the set of d-dimensional Hermitian matrices $\mathsf{H}(d)$ so that an arbitrary $\rho \in \mathsf{D}(d)$ can be written

$$\rho = \frac{1}{d}\sum_{i=1}^{d^2} \mathrm{Tr}(P_i\rho)P_i.$$

The straightforward algorithm here is then to estimate each of the coefficients $\mathrm{Tr}(P_i\rho)$ with sufficient accuracy, which will serve as a complete description of the estimate of ρ. Consider the d^2 POVMs $\mathcal{M}_i$ with corresponding measurement operators $\{\frac{1}{2}(\mathbb{1} \pm P_i)\}$ for each $i \in [d^2]$, with possible outcomes $z_i \in \{\pm 1\}$ defined in the obvious way. Then z_i is an unbiased estimator for the i^{th} Pauli coefficient, and performing this measurement $s \in \mathbb{Z}_+$ times results in iid random variables $\{z_{i,j}\}_{j=1}^s$. Let us then take the empirical average of the s samples corresponding to the i^{th} Pauli measurement $\boldsymbol{\mu}_i := \frac{1}{s}\sum_{j=1}^s z_{i,j}$, for each $i \in [d^2]$, which requires a total of sd^2 measurements on separate copies of ρ. We then consider our estimate of the state to be $\hat{\boldsymbol{\rho}} := \frac{1}{d}\sum_{i=1}^{d^2} \boldsymbol{\mu}_i P_i$, which clearly satisfies $\mathbb{E}\,\hat{\boldsymbol{\rho}} = \rho$. We may then compute

$$\mathbb{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_{\mathrm{F}}^2 = \frac{1}{d}\sum_{i=1}^{d^2} \mathbb{E}\,|\boldsymbol{\mu}_i - \mathrm{Tr}(P_i\rho)|^2$$

$$
\begin{aligned}
&= \frac{1}{d} \sum_{i=1}^{d^2} \mathrm{Var}[\boldsymbol{\mu}_i] \\
&= \frac{1}{ds^2} \sum_{i=1}^{d^2} \sum_{j=1}^{s} \mathrm{Var}[\boldsymbol{z}_{i,j}] \\
&\leq \frac{d}{s}
\end{aligned}
$$

where in the third line we used the property $\mathrm{Var}[a\boldsymbol{x}] = a^2 \mathrm{Var}[\boldsymbol{x}]$ for a random variable $\boldsymbol{x}$, as well as the fact that the variance is additive for independent random variables. The final line follows since $|\boldsymbol{z}_{i,j}| = 1$. Using the inequality $\|\cdot\|_1 \leq \sqrt{d}\,\|\cdot\|_{\mathrm{F}}$, we find for $s = d^2/\epsilon^2$, it holds that $\mathbb{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_1 \leq \epsilon$. We can once again convert this statement about convergence in expectation to convergence with high probability using Markov's inequality, which leads to the conclusion that ϵ-accurate tomography in trace distance is achievable using at most $sd^2 = d^4/\epsilon^2$ binary Pauli measurements on separate copies of ρ.

C Measurements with polynomial-size circuits

Theorems 5.6, 5.7, and 6.3 give lower bounds for quantum learning tasks in the setting of adaptive measurements, when they are drawn from a finite set of possible measurements. These results thus also limit the power of adaptivity using measurements that can be implemented with polynomial-size circuits. This includes efficiently implementable measurements, i.e., measurements whose circuits are also uniformly generated. In this section, we explain what it means for a family of measurements to have polynomial-size circuits. Fix a (possibly infinite) universal gate set $\mathcal{G}$ consisting of constant-arity gates (e.g., one- and two-qubit gates).

Definition C.1 (Measurements with polynomial-size circuits, constant number of outcomes)**.** For any $q \geq 1$, suppose $\mathcal{A}_q$ is a collection of measurements acting on q-qubit quantum states. We say the family of measurements $(\mathcal{A}_q : q \geq 1)$ has *polynomial-size* if there exist polynomials p_1, p_2 such that for each q and measurement $\mathcal{M} \in \mathcal{A}_q$ there is a quantum circuit on $q + p_1(q)$ qubits with at most $p_2(q)$ gates from $\mathcal{G}$ that implements $\mathcal{M}$. I.e., the measurement $\mathcal{M}$ has the action

$$
\mathcal{M} : \rho \mapsto \sum_{y \in \{0,1\}^S} \langle y| \operatorname{Tr}_{[\ell]\setminus S} \left(U(\rho \otimes |\overline{0}\rangle\langle\overline{0}|)U^\dagger \right) |y\rangle \, |y\rangle\langle y|
$$

for any state $\rho \in \mathsf{D}(2^q)$, where $S \subseteq [\ell]$, $\ell := q + p_1(q)$, $|\overline{0}\rangle \in (\mathbb{C}^2)^{\otimes p_1(q)}$ is the all-zero state for $p_1(q)$ ancilla qubits and $U \in \mathsf{U}(2^{q+p_1(q)})$ is the unitary operator given by the composition of the gates in the circuit.

We say that the measurements in the family have a constant number of outcomes if there is a positive integer r such that for all measurements $\mathcal{M} \in \mathcal{A}_q$ and for all $q \geq 1$, we have $\mathrm{rank}(\mathcal{M}) \leq r$.

In the case where $\mathcal{G}$ is finite, by a counting argument we may verify that the number of distinct measurements m in $\mathcal{A}_q$ for any family with polynomial size is at most $\mathrm{poly}(q)^{\mathrm{poly}(q)}$ which is in $\exp(o(d))$, where $d := 2^q$ is the dimension of the system. It follows immediately from Theorem 5.6 and this bound that $\Omega(d^3/\epsilon^2)$ single-copy, possibly adaptive, efficient measurements are necessary to perform tomography. (Note that *efficient* measurements are also required to be *uniformly generated*, in addition to having polynomial-size circuits.) Similarly, we may infer a bound for shadow tomography using only efficient single-copy measurements from Theorem 6.3 (see the remark after the theorem).

We may extend this reasoning to the case where $\mathcal{G}$ has infinite cardinality, but consists of gates of constant arity — for example, when all single-qubit gates are included in the set. This comes at the cost of the loss of a multiplicative factor of at most $\mathrm{polylog}(1/\epsilon)$ in the lower bounds. This is accomplished by an application of the Solovay-Kitaev theorem and adjusting the general argument we have been using to prove the bounds. We replace each measurement with a suitably accurate approximation with a circuit over a finite gate set, and show that the approximation results in at most a small constant deviation from the original distribution over measurement outcomes. In the sequel, we refer to the case where a learner performs measurements with circuits over the gate set $\mathcal{G}$ as the *original* strategy. Fix any *finite* universal gate set $\mathcal{G}'$ that contains the inverses of all the gates in it.

Proposition C.2. *Let $d := 2^q$ for some $q \geq 1$. Suppose a learner performs $n \in O(d^3/\epsilon^2)$ adaptive measurements on single copies of quantum states in $\mathsf{D}(d)$, where each measurement can be implemented with a circuit of size at most a polynomial t in q using an infinite set $\mathcal{G}$ of gates with constant arity. There is an adaptive measurement strategy consisting of single-copy measurements with circuits of size of order $qt(\log q + \log(1/\epsilon))$ over $\mathcal{G}'$ such that for any state $\rho \in \mathsf{D}(d)$, the distribution over the n measurement outcomes obtained from measuring ρ is 0.01-close in total variation distance to the corresponding distribution obtained with the original strategy.*

Proof. Suppose learner performs measurements with circuits of size at most t over the gate set $\mathcal{G}$ in the original strategy. Suppose that this learner obtains the outcomes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ using the original strategy, and consider the i^{th} measurement in the sequence $\mathcal{M}_i^{y_{<i}} : \mathsf{D}(2^q) \to \mathsf{D}(2^{u(q)})$ for some fixed sequence of previous outcomes $y_{<i}$, and polynomial $u(q)$. By the Solovay-Kitaev Theorem, for any $\delta \in (0,1)$ there is a measurement $\Phi_i^{y_{<i}} : \mathsf{D}(2^q) \to \mathsf{D}(2^{u(q)})$ which can be implemented using circuits of size $t' := t \cdot \mathrm{polylog}(t/\delta)$ gates from $\mathcal{G}'$ and which satisfies

$$\left\| \mathcal{M}_i^{y_{<i}} - \Phi_i^{y_{<i}} \right\|_\diamond \leq 2\delta \ , \tag{38}$$

where $\|\cdot\|_\diamond$ is the completely bounded trace norm (some times also called the diamond norm). In other words, for any fixed state $\rho \in \mathsf{D}(2^q)$ the total variation distance between the distributions over outcomes obtained by measuring ρ according to the two measurements is at most δ.

Suppose the learner adopts the *modified strategy* given by the measurements $\Phi_i^{y_{<i}}$ for every $i \in [n]$, given the previously observed outcomes $y_{<i}$. We show by induction that the deviation of the resulting distribution from that of the original strategy grows linearly with n. For each $i \in [n]$, let $\boldsymbol{y}'_i$ denote the measurement outcome from the i^{th} measurement using the modified strategy. Note that these random variables have the same set of possible outcomes, which are bit-strings of length at most $\mathrm{poly}(q)$. Define the corresponding conditional distributions p and ϕ over outcomes as

$$p(y_k|y_{<k}) := \Pr\left[\boldsymbol{y}_k = y_k | \boldsymbol{y}_{<k} = y_{<k}\right], \qquad \phi(y_k|y_{<k}) = \Pr\left[\boldsymbol{y}'_k = y_k | \boldsymbol{y}'_{<k} = y_{<k}\right]$$

as well as marginal probabilities $p(y_{<k})$, $\phi(y_{<k})$, for each $k \in [1,n]$. Let us also define the notation $y_{\leq k} := y_{<k+1}$. For the first measurement outcome, the total variation distance between the two distributions is $\frac{1}{2}\sum_{y_1} |p(y_1) - \phi(y_1)| \leq \delta$, using Eq. (38). Now suppose that

$$\sum_{y_{<k}} |p(y_{<k}) - \phi(y_{<k})| \leq 2(k-1)\delta.$$

for some $k > 1$. Then we have

$$\sum_{y_{\leq k}} |p(y_{\leq k}) - \phi(y_{\leq k})| = \sum_{y_k} \sum_{y_{<k}} |p(y_k|y_{<k})p(y_{<k}) - \phi(y_k|y_{<k})\phi(y_{<k})|$$

$$
\begin{aligned}
&\leq \sum_{y_k}\sum_{y_{<k}} |p(y_k|y_{<k})p(y_{<k}) - \phi(y_k|y_{<k})p(y_{<k})| \\
&\qquad + \sum_{y_k}\sum_{y_{<k}} |\phi(y_k|y_{<k})p(y_{<k}) - \phi(y_k|y_{<k})\phi(y_{<k})| \\
&= \sum_{y_{<k}} p(y_{<k}) \sum_{y_k} |p(y_k|y_{<k}) - \phi(y_k|y_{<k})| + \sum_{y_{<k}} |p(y_{<k}) - \phi(y_{<k})| \\
&\leq 2\delta + 2(k-1)\delta \\
&= 2k\delta.
\end{aligned}
$$

Hence, the total variation distance between the two distributions corresponding to all n outcomes is at most $n\delta$. Taking $\delta = 1/(100n)$ ensures that the total error from the modified strategy is at most 0.01. Moreover, all measurements in the modified strategy are implemented with circuits of size at most $t \cdot \mathrm{polylog}(100nt)$ over the finite gate set $\mathcal{G}'$. Since $t(q)$ is a polynomial in q and $n \in O(d^3/\epsilon^2)$, the total number of gates t' is of order $qt(\log q + \log(1/\epsilon))$. □

Note that the total variation distance being equal to 0.01 is not significant — the point is that this modified strategy only affects the success probability of the learning procedure by a small constant. For example, consider the task of quantum state tomography with adaptive single-copy measurements. Let $\boldsymbol{x}$ be distributed over $\mathsf{D}(d)$ and $\boldsymbol{y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be the measurement outcomes obtained using the original strategy on $n \in O(d^3/\epsilon^2)$ copies of $\boldsymbol{x}$. Let $\boldsymbol{y}' = \boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_n$ be the outcomes obtained using the modified strategy on n copies of $\boldsymbol{x}$. Prop. C.2 says that for any value x of the random variable $\boldsymbol{x}$,

$$
\left\| p_{\boldsymbol{y}|x} - p_{\boldsymbol{y}'|x} \right\|_1 \leq 1/100 \ .
$$

Therefore, if the original strategy succeeds in identifying the state to within accuracy ϵ with "high" probability (say, $\geq 2/3$) then the modified strategy succeeds with high probability as well. However, the modified strategy uses measurements drawn from a finite set of measurements, which is the setting for which our lower bounds apply. By counting the number of distinct circuits of size t' we see that the total number of distinct measurements m used in the modified strategy is at most $\mathrm{poly}(t')^{t'}$. So $\log m$ is of the order of

$$
\mathrm{poly}(q)(\log q + \log(1/\epsilon)) \ .
$$

By Theorem 5.6 and Prop. C.2 we have that

$$
\Omega\left(\frac{d^3}{\epsilon^2(1 + \mathrm{polylog}(d)\, u(d, 1/\epsilon)/d)}\right) \tag{39}
$$

samples are required, where

$$
u(d, 1/\epsilon) \coloneqq (\log\log d + \log(1/\epsilon)) \log(\log\log d + \log(1/\epsilon)) \ .
$$

Similarly, for the single-copy shadow tomography problem (cf. Def. 6.2), by Theorem 6.3

$$
\Omega\left(\frac{d \min\{\log(M),\, d^2\}}{\epsilon^2(1 + \mathrm{polylog}(d)\, u(d, 1/\epsilon)/d)}\right) \tag{40}
$$

samples are required, even when the measurements are implemented efficiently using a constant arity gate set of possibly infinite cardinality. Note that these bounds are asymptotically smaller than those for finite gate sets only when the approximation parameter ϵ is exponentially small in the dimension d.