Sample-optimal tomography of quantum states

Jeongwan Haah,[1,2] Aram W. Harrow,[2] Zhengfeng Ji,[3,4,5] Xiaodi Wu,[6] and Nengkun Yu[4,3,7]

[1] *Station Q Quantum Architectures and Computation, Microsoft Research, Redmond, Washington, USA*
[2] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*
[3] *Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, NSW 2007, Australia*
[4] *Institute for Quantum Computing, University of Waterloo, Waterloo, Ontario, Canada*
[5] *State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China.*
[6] *Department of Computer and Information Science, University of Oregon, Eugene, Oregon, USA*
[7] *Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, Canada*

It is a fundamental problem to decide how many copies of an unknown mixed quantum state are necessary and sufficient to determine the state. Previously, it was known only that estimating states to error ϵ in trace distance required $O(dr^2/\epsilon^2)$ copies for a d-dimensional density matrix of rank r. Here, we give a theoretical measurement scheme (POVM) that requires $O(dr/\delta)\ln(d/\delta)$ copies of ρ to error δ in infidelity, and a matching lower bound up to logarithmic factors. This implies $O((dr/\epsilon^2)\ln(d/\epsilon))$ copies suffice to achieve error ϵ in trace distance. We also prove that for independent (product) measurements, $\Omega(dr^2/\delta^2)/\ln(1/\delta)$ copies are necessary in order to achieve error δ in infidelity. For fixed d, our measurement can be implemented on a quantum computer in time polynomial in n.

Given n copies of an unknown d-dimensional quantum state ρ, how accurately can ρ be estimated? This fundamental question arises both in quantum information theory and in the interpretation of experimental results. Since ρ has d^2-1 real parameters, it is reasonable to conjecture that $\Theta(d^2)$ measurements are necessary and sufficient to estimate ρ to constant accuracy. On the other hand, even distinguishing a fair coin from a coin biased to obtain heads with probability $1/2+\epsilon$ requires $\Omega(1/\epsilon^2)$ measurements.

In this paper we show that the number of copies required to estimate ρ with precision ϵ scales roughly with both d^2 and $1/\epsilon^2$. More precisely, if the fidelity goal is $1-\delta$, we prove an $\Omega(d^2/\delta)$ lower bound and an $O((d^2/\delta)\ln(d/\delta))$ upper bound on the number of required copies. When the state ρ is guaranteed to have rank $\le r$ we show an $O((dr/\delta)\ln(d/\delta))$ upper bound and an $\Omega((dr/\delta)/\ln(d/r\delta))$ lower bound. We also prove a lower bound $\Omega(dr^2/\delta^2)/\ln(1/\delta)$ for independent measurement schemes where individual copies are measured independently and then the outcomes are processed to output an estimate $\hat{\rho}$. Our result is summarized in Table I.

Notation We use the convention that $\Omega(x)$ means a function that is asymptotically $\ge c_1 x$ for a constant $c_1>0$, $O(x)$ means $\le c_2 x$ for a constant $c_2>0$ and $\Theta(x)$ means both $O(x)$ and $\Omega(x)$. Notation $\tilde{O}()$ means that we neglect ln factors. ln and exp are base-e.

I. ACCURACY MEASURES

The fidelity of two quantum states ρ,σ is $F(\rho,\sigma) := \mathrm{tr}\sqrt{\sqrt{\rho}\,\sigma\sqrt{\rho}}$, the "infidelity" is $1-F$, represented by δ, and their trace distance is $T(\rho,\sigma) := \frac{1}{2}\|\rho-\sigma\|_1$, represented by ϵ. These are related by [3]

$$1-F \le T \le \sqrt{1-F^2}. \tag{1}$$

We derive an upper bound in terms of fidelity and a lower bound in terms of trace distance, in each case implying a near-optimal bound in terms of the other quantity. Here we discuss why fidelity is in many ways a natural quantity for tomography [4]. Tomography is essentially a state discrimination procedure where one distinguishes $\rho^{\otimes n}$ from $\sigma^{\otimes n}$. The statistical distinguishability of these states is measured by the trace distance $T_n = T(\rho^{\otimes n},\sigma^{\otimes n})$, which is in general much larger than $T(\rho,\sigma)$; this amplification is what enables the tomography. The asymptotic behavior of T_n can be quantified as

$$\frac{1}{2}F(\rho,\sigma)^{2n} \le 1-T_n \le F(\rho,\sigma)^n$$

by Eq. (1) and $F(\rho^{\otimes n},\sigma^{\otimes n}) = F(\rho,\sigma)^n$. This means that $\ln(1/F)$ or infidelity gives nearly sharp bounds on the rate at which T_n converges to 1; the actual rate[1] is between $\ln(1/F)$ and $2\ln(1/F)$. In particular, for fixed d, the state discrimination is possible to infidelity δ using $n=\Theta(1/\delta)$ copies. Our upper bound on n in terms of fidelity proves that the POVM we present in this paper indeed accomplishes the discrimination task using $n=\tilde{O}(1/\delta)$ copies. On the contrary, the corollary upper bound in terms of trace distance sometimes over-estimates the

[1] The exact scaling of $1-T_n$ for large n is known to be C^n where $C = C(\rho,\sigma) = \inf_{0\le s\le 1}\mathrm{tr}(\rho^s\sigma^{1-s})$, and $\ln(1/C)$ is called the quantum Chernoff distance [5, 6].

TABLE I. Conditions for the quantum state tomography with high success probability. δ denotes the accuracy goal measured in the infidelity $1 - F(\rho, \hat{\rho}) = 1 - \|\sqrt{\rho}\sqrt{\hat{\rho}}\|_1$, and ϵ denotes that in the trace distance $T(\rho, \hat{\rho}) = \frac{1}{2}\|\rho - \hat{\rho}\|_1$. The upper bound in terms of the infidelity implies that in terms of trace distance; $n \leq O(d^2/\epsilon^2)\ln(d/\epsilon)$. The lower bound in terms of the trace distance implies that in terms of infidelity; e.g. $n \geq \Omega(d^2/\delta)$. The lower bound for the independent measurements in rank r case implies $n \geq \Omega(dr^2/\epsilon^2 \ln(1/\epsilon))$. The previously known upper bound on n already used only independent measurements; thus our lower bounds show that this result was essentially optimal.

	Our result		Previous result
	for general $\rho \in \mathbb{C}^{d\times d}$	for ρ of rank at most r	
Sufficient	$n \leq O(d^2/\delta)\ln(d/\delta)$	$n \leq O(rd/\delta)\ln(d/\delta)$	$n \leq O(r^2 d/\epsilon^2)$ [1] See Sec. II A.
Necessary	$n \geq \Omega\left(d^2/\epsilon^2\right)$	$n \geq \Omega\left(rd/\epsilon^2\right)/\ln(d/r\epsilon)$	$n \geq \Omega(1/\epsilon^2) + \tilde{\Omega}(rd)$ [2]
Necessary using independent measurements	$n \geq \Omega(d^3/\epsilon^2)$	$n \geq \Omega(dr^2/\delta^2 \ln(1/\delta))$	$n \geq \Omega(1/\delta^2 \ln(1/\delta))$ See Sec. II.

sufficient number of samples by an unbounded amount. As a simple example, consider qubit states

$$\rho = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \sigma = \begin{pmatrix} 1-\epsilon & 0 \\ 0 & \epsilon \end{pmatrix},$$

between which the trace distance is ϵ and the infidelity is $1 - \sqrt{1-\epsilon} \simeq \epsilon/2$. The trace distance bound only says $n = \tilde{O}(1/\epsilon^2)$ copies are sufficient to distinguish them, whereas the fidelity bound says $n = \tilde{O}(1/\epsilon)$ copies are sufficient.

II. PREVIOUS RESULTS

Quantum state estimation has been extensively studied, going back at least to the work of Helstrom [7], Holevo [8] and others from around 1970. Many of the rigorous results are for the special cases when $d = 2$ or $r = 1$, or give an uncontrolled or suboptimal d dependence (e.g. with n scaling as $f(d)/\delta$ for unknown f) or discuss related problems such as spectrum estimation, parameter estimation or determining the identity of a state drawn from a discrete set. In this paper we will consider optimal measurements (also called "collective" measurements) and will not discuss the extensive literature on independent or adaptive measurements.

For $d = 2$ (i.e. qubits), the optimal infidelity was shown in [9–13] to scale as $1/n$. This scaling was generalized to qudits in [14] (see also Section 6.4 of [15]), but with an uncontrolled dependence on d (i.e.n scales as $f(d)/\delta$ for unknown $f(\cdot)$); see also [16]. In many settings (e.g. minimax estimation) one can show that covariant measurements are optimal. If one further assumes that ρ is pure then the optimal estimation strategy has a simple form and n should scale as $\Theta(d/\delta)$ [8, 17]; see also [18] where further connections were made to cloning and de Finetti theorems.

Another major theme in recent work has been the study of various forms of restricted measurements, e.g. independent measurements with a limited number of measurement settings. Intermediate between independent measurements and unrestricted (also called "collective" or "entangled") measurements are *adaptive* measurements in which the copies of ρ are measured individually, but the choice of measurement basis can change in response to earlier measurements.

On the achievability side for independent measurements, a sequence of works [1, 19–21] showed that $n = O(dr^2/\epsilon^2)$ copies are sufficient to obtain trace distance $\leq \epsilon$ with high probability.[2] On the other hand, even for $d = 2$, adaptive and collective measurements are known to have asymptotically better error scaling, at least when measured in terms of infidelity. The usual intuition is that n should scale as $1/\delta^2$ for independent measurements and $1/\delta$ for adaptive or collective measurements; e.g. see [22] for numerical evidence. Refs. [10, 13] showed that adaptive measurements could achieve $n = O(1/\delta)$ scaling. When a POVM contains a finitely many elements, the lower bound $1/\delta^2$ can be demonstrated by considering qubit tomography when the density matrix does not commute with POVM elements. We were unable to find a reference that proves this particular fact. Ref. [23] gave an $\Omega(\frac{1}{\delta'^2 \ln(1/\delta')})$ lower bound for independent measurements with *relative entropy* δ' as accuracy measure without restriction that POVM should consist of finitely many elements.

In many cases it is not necessary to determine the full state ρ but only to estimate some parameters of the state. This is an extremely general problem which includes results such as a quantum version of the Cramér-Rao bound [7, 24, 25] again going back to the early prehistory of quantum information. One special case that uses similar representation-theory techniques to our work is the problem of spectrum estimation. Here, the optimal covariant measurement was described by Keyl and Werner [26], its large-deviation properties were derived in [27] (see also [28]), and it was analyzed further in [29, 30]. Ref. [30] in particular showed (among other results) that

[2] The earlier papers [19, 20] achieved $n = \tilde{O}(d^2 r^2/\epsilon^2)$. The improved $n = O(dr^2/\epsilon^2)$ performance is achieved by analyzing Theorem 2 of [1]. This is not obvious from their theorem statement, but we explain the connection in Sec. II A.

the Keyl-Werner algorithm required

$$\Omega\left(\frac{d^2}{\epsilon^2}\right) \leq n \leq O\left(\frac{d^2}{\epsilon^2}\ln\frac{d}{\epsilon}\right).$$

Our results improve the upper bound by using the same number of copies to obtain a full estimate of ρ instead of merely its spectrum. We also improve the lower bound by showing that it applies to *all* estimation strategies, not only the Keyl-Werner algorithm; on the other hand, our lower bound is for the harder problem of state estimation, while the lower bound of Ref. [30] is for the problem of spectrum estimation. We improve both bounds in the case when $r \ll d$.

The problem of quantum state estimation can be thought of as a special case of minimax estimation (i.e. choosing an estimator that minimizes the expected loss when we maximize over input states) when the loss function is given by the infidelity. Other loss functions have also been considered [31, 32]. For example, with the 0-1 loss function (assuming ρ is drawn from a finite set) the goal is to maximize the probability of guessing ρ correctly. Here a powerful heuristic is to use the so-called "pretty good measurement" or PGM [33–35], whose error is never worse than twice that of the optimal measurement for any ensemble [36]. While the PGM requires a prior distribution, prior-free versions can also be constructed [37]. We will describe two closely related measurements in this paper: first, one closely related to the PGM and then one (with roughly equivalent performance) that corresponds precisely to a PGM over an appropriately chosen "uniform" ensemble of density matrices. In each case, we analyze the measurements directly, without making use of the results of [36, 37] or other prior work.

A. Sample complexity in Kueng et al. [1]

The previously best achievable sample complexity for state tomography was described in [1]. Their setting does not naturally translate into our framework, so for convenience we sketch here how that is achievable. First we restate one of their main theorems:

Theorem 1. *There are universal constants $C_1, C_2, C_3 > 0$ such that the following holds for any r, d. Let $a_1, \ldots, a_m \in \mathbb{C}^d$ be independent standard Gaussian vectors; i.e. normalized such that $\mathbb{E}[|a_i\rangle\langle a_j|] = I_d\delta_{ij}$. If $m \geq C_1 dr$, then with probability $\geq 1 - e^{-C_2 m}$ our choice of $a_1, \ldots, a_m$ is "good" in a sense we will define below.*

For X a matrix, define $\mathcal{A}(X) = \sum_j \langle a_j| X |a_j\rangle |j\rangle \in \mathbb{R}^m$. Given a d-dimensional density matrix ρ, a vector $b \in \mathbb{R}^m$ and a noise parameter η, define σ be any minimum of the following convex program:

$$\min \|\sigma\|_1 \text{ subject to } \|\mathcal{A}(\sigma) - b\|_2 \leq \eta.$$

Suppose further that $\|\mathcal{A}(\rho) - b\|_2 \leq \eta$. If the vectors $a_1, \ldots, a_m$ are good, then we have

$$\|\rho - \sigma\|_2 \leq C_3 \frac{\eta}{\sqrt{m}}. \tag{2}$$

To translate this into a quantum measurement, observe that by the operator Chernoff bound [38], we have $\frac{1}{m}\sum_{i=1}^m |a_i\rangle\langle a_i| \approx I_d$ with high probability. (For the purpose of this analysis, we neglect the error here.) We can then define a POVM with elements $E_i = |a_i\rangle\langle a_i|/m$. Measuring this POVM yields outcome i with probability $p_i := \mathrm{tr}[E_i\rho]$; in the notation of [1] we have $p = \mathcal{A}(\rho)/m$. We will define the vector b of observed probabilities by measuring n independent copies of ρ using this POVM. If the resulting vector of frequencies is f, i.e., outcome i occurs f_i times, then we define $b = \frac{m}{n}f$. Thus b is an unbiased estimator of $\mathcal{A}(\rho)$; i.e. $\mathbb{E}[b] = \frac{m}{n}\mathbb{E}[f] = \frac{m}{n}np = \mathcal{A}(\rho)$. We can also estimate the error by

$$\mathbb{E}\,\|b - \mathbb{E}[b]\|_2^2 = \frac{m^2}{n^2}\sum_{i=1}^m \mathrm{Var}[f_i] \leq \frac{m^2}{n^2}\sum_{i=1}^m np_i = \frac{m^2}{n}.$$

We thus have $\eta \leq O(m/\sqrt{n})$ with high probability. According to (2) we then have $\|\rho - \sigma\|_2 \leq O(\sqrt{m/n}) = O(\sqrt{dr/n})$. It follows that

$$\begin{aligned}&\|\rho - \sigma\|_1\\ &\leq 2\sqrt{\min(\mathrm{rank}(\rho), \mathrm{rank}(\sigma))}\|\rho - \sigma\|_2\\ &\leq O(\sqrt{dr^2/n}).\end{aligned}$$

In other words, trace-distance error ϵ can be achieved with $n = O(dr^2/\epsilon^2)$. While this bound is significantly worse than our bound of $\tilde{O}(dr/\epsilon^2)$, their approach does have the significant advantage of not requiring entangled measurements. The improved performance of our bound (as well as that of [39]) can be seen as the advantage that entangled measurements yield for tomography.

B. Two-stage measurement scheme using local asymptotic normality

The local asymptotic normality in Ref. [12, 40] asserts that n copies $\rho_\theta^{\otimes n}$ of states ρ_θ in a sufficiently small neighborhood $\mathcal{B}$ of a state $\rho_{\theta=0}$ behaves like an ensemble of Gaussian states of quantum harmonic oscillators. In relation to our discussion of state estimation, it is important that there exists a channel [12, 40], which is faithful in the limit $n \to \infty$, from $\rho_\theta^{\otimes n}$ to gaussian states, such that one can estimate the parameter θ of the state optimally. The size of the neighborhood in the correspondence in fact depends on n. Theorem 4.1 in Ref. [40] gives a lower bound on this size, which reads

$$\mathcal{B} \supseteq \{\rho \;:\; \|\rho - \rho_0\|_2 \leq n^{-1/2+\eta}\}$$

where $\eta \in (0, 1/6)$.

Based on this result, Ref. [40] proposes a two-stage adaptive measurement scheme of a completely unknown

state. In the first stage, using n_1 copies of the state, one "roughly" measures the state in order to have a confidence region inside $\mathcal{B}$. The actual measurement method for this first stage is not shown, and we assume that this is a non-adaptive independent measurement on each copy. In the second stage, one uses remaining $n_2 = n - n_1$ copies and apply the local asymptotic normality to optimally estimate the state.

Let us analyze the sample complexity of this proposal. We assume that the channel between $\rho_\theta^{\otimes n}$ and Gaussian states is exactly faithful for any n. This assumption may not be true on its own, but is certainly a favorable condition to assess the advantage of local asymptotic normality. After the first stage, the size of the confidence region must be $\epsilon_i = n_2^{-1/2+\eta}$ in 2-norm. This requires at least $n_1 \geq \Omega(d^2/\epsilon_i^2)$, by our lower bound Theorem 4. Suppose ϵ_f is our accuracy goal in 2-norm. If $\epsilon_f \geq \epsilon_i$, then the second stage becomes redundant, and overall measurement is by the non-adaptive independent measurement. Our result says that this scheme cannot be sample-optimal. If $\epsilon_f < \epsilon_i$, then one needs $n_2 \geq \Omega(d/\epsilon_f^2)$ in the second stage. To achieve ϵ-accuracy in 1-norm, we must have $\epsilon_f \leq \epsilon/\sqrt{d}$, and overall sample complexity becomes

$$n = n_1 + n_2 \geq \Omega(d^{4-4\eta}/\epsilon^{2-4\eta}) + \Omega(d^2/\epsilon^2).$$

Since $\eta < 1/6$, the dependence of n on d is actually worse than the independent non-adaptive scheme, although the dependence of n on ϵ is optimal. In other words, the measurement scheme using the asymptotic normality may yield asymptotically optimal error scaling, but it takes too many samples to enter the regime where the asymptotic normality becomes useful for high dimensional states.

III. REVIEW ON REPRESENTATION THEORY OF UNITARY AND SYMMETRIC GROUPS

Schur-Weyl duality is a statement regarding joint representations of a matrix group and the symmetric group. This is standard material [41] in representation theory, but for the reader's convenience we explain parts that are relevant to our results.

Consider the Hilbert space $\mathcal{H} = (\mathbb{C}^d)^{\otimes n}$ of n qudits of d-dimensions. This space admits representations of the general linear group $GL(d)$ and the symmetric group $\mathbb{S}_n$. The matrix group acts by simultaneous "rotation" as $U^{\otimes n}$ for any $U \in GL(d)$, and the symmetric group acts by permuting tensor factors. Concretely, a permutation $\pi \in \mathbb{S}_n$ is represented by

$$P_\pi = \sum_{\{j_i\}} \left|j_{\pi^{-1}(1)} j_{\pi^{-1}(2)} \cdots j_{\pi^{-1}(n)}\right\rangle \left\langle j_1 j_2 \cdots j_n\right|.$$

Two actions $U^{\otimes n}$ and P_π obviously commute with each other, and hence $\mathcal{H}$ admits a representation of $G = GL(d) \times \mathbb{S}_n$. The Schur-Weyl duality states that these two representations are commutants of each other on $\mathcal{H}$. That is, if a matrix K on $\mathcal{H}$ commutes with all P_π, then $K = \sum_i c_i U_i^{\otimes n}$ for some $U_i \in GL(d)$ and numbers $c_i \in \mathbb{C}$. Conversely, if a matrix K on $\mathcal{H}$ commutes with all $U^{\otimes n}$, then $K = \sum_\pi c_\pi P_\pi$ for some $c_\pi \in \mathbb{C}$.

Generally, an irreducible representation (irrep) of G is given by the tensor product of an irrep of $GL(d)$ and an irrep of $\mathbb{S}_n$. Since the two groups are mutual commutants on $\mathcal{H}$, the irreps in $\mathcal{H}$ of the two groups must be in a one-to-one correspondence. They are specified by Young diagrams, or equivalently, partitions $\lambda = (\lambda_1, \ldots, \lambda_n)$ of $n = \sum_i \lambda_i$, where λ is sorted to be non-increasing. Thus, we have a decomposition

$$(\mathbb{C}^d)^{\otimes n} = \bigoplus_{\lambda \vdash n} \Pi_\lambda (\mathbb{C}^d)^{\otimes n} = \bigoplus_{\lambda \vdash n} \mathcal{Q}_\lambda \otimes \mathcal{P}_\lambda$$

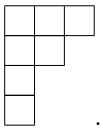
where $\mathcal{Q}_\lambda$ is the irrep of $\mathrm{GL}(d)$ and $\mathcal{P}_\lambda$ is the irrep of $\mathbb{S}_n$, and Π_λ is the projector onto the component $\mathcal{Q}_\lambda \otimes \mathcal{P}_\lambda$. Direct consequences of the decomposition are that

$$\Pi_\lambda X^{\otimes n} \Pi_\lambda \cong \mathbf{q}_\lambda(X) \otimes \mathrm{id}_{\mathcal{P}_\lambda} \tag{3}$$

$$\Pi_\lambda X^{\otimes n} = X^{\otimes n} \Pi_\lambda \tag{4}$$

for *any* $d \times d$ matrix X, where we have defined $\mathbf{q}_\lambda(X)$ to mean the representing matrix of X. In fact, this is the main reason we are dealing with $GL(d)$, which is dense in the set of all matrices, rather than the more familiar $\mathbb{U}(d)$. The space $\mathcal{Q}_\lambda$ is also an irrep of the unitary group $\mathbb{U}(d)$, and our discussion of Schur-Weyl duality could have been formulated entirely with $\mathbb{U}(d)$; however, under this formulation X would be restricted to be unitary.

For our results it is important to understand the characters of the irrep $\mathcal{Q}_\lambda$ of $GL(d)$. We identify a partition λ with a *Young diagram* in which there are λ_i boxes in the i^{th} row, e.g. the diagram for $\lambda = (3, 2, 1, 1)$ is as follows



Define a Young tableau T with shape λ to be a way of filling each box in λ with a number, e.g.

5	5	2
1	2	3
4		
1		

A *standard Young tableau* (SYT) is one in which each number from $1, \ldots, n$ appears exactly once and numbers strictly increase from left to right and from top to bottom, while in a semi-standard Young tableau (SSYT) numbers weakly increase from left to right and strictly increase from top to bottom. Associated with a standard Young tableau T there are two subgroups A_T and B_T of $\mathbb{S}_n$. A_T is the set of all permutations that permute numbers within the rows of T, and B_T is the set of all permutations that permute numbers within the columns of T. The Young symmetrizer is then defined as

$$Y_T = \sum_{a \in A_T, b \in B_T} \mathrm{sgn}(b) P_a P_b.$$

It can be shown that Y_T is proportional to an orthogonal projector, and it turns out that $Y_T\mathcal{H}$ is an irrep of $GL(d)$ and is isomorphic to $\mathcal{Q}_\lambda$. Since every T with the same λ gives rise to an isomorphic irrep of $GL(d)$, let us set T to be the SYT where $1, 2, \ldots, n$ are written in order from the upper left box towards right and down. To understand the basis of $\mathcal{Q}_\lambda$, let $|1\rangle, |2\rangle, \ldots, |d\rangle$ form the standard orthonormal basis of $\mathbb{C}^d$. We may regard each basis vector $|E\rangle = |j_1, \ldots, j_n\rangle$ of $\mathcal{H}$ as a Young tableau E of shape λ. The Young symmetrizer Y_T projects this basis vector to a vector of $\mathcal{Q}_\lambda$. If there is any repetition along a column of E, then Y_T will annihilate it, thanks to the antisymmetric sum over P_b for $b \in B_T$. It follows that $\mathcal{Q}_\lambda = 0$ whenever λ has more than d rows. More precisely, let $\nu_i = \nu_i(E)$ denote the number of times the basis element $|i\rangle$ appears in the tableau E (also known as the *weight* of E), and let $\nu^\downarrow$ be the vector obtained by sorting ν into non-increasing order. Then Y_T annihilates E whenever $\sum_{i=1}^m \nu_i^\downarrow > \sum_{i=1}^m \lambda_i$ for some $m = 1, \ldots, d-1$. The negation of the last condition is often denoted as

$$\nu \prec \lambda \Leftrightarrow \begin{cases} \sum_{i=1}^m \nu_i^\downarrow \le \sum_{i=1}^m \lambda_i & (1 \le m < d) \\ \sum_{i=1}^d \nu_i^\downarrow = \sum_{i=1}^d \lambda_i \end{cases}$$

and we say that ν *is majorized by* λ. The surviving tableaux E with $\nu(E) \prec \lambda$ form a spanning set for $\mathcal{Q}_\lambda$, or if we restrict to SSYT, they form a basis.

Now we can derive an expression for the characters of $\mathcal{Q}_\lambda$. Since $\operatorname{tr} \mathbf{q}_\lambda(X)$ must be a function of eigenvalues of X, we may assume without loss of generality that X is a diagonal matrix with eigenvalues $x_1, \ldots, x_d$ associated with the standard basis elements $|1\rangle, \ldots, |d\rangle$. The basis vectors of $\mathcal{Q}_\lambda$ we just constructed are eigenvectors of diagonal $X^{\otimes n}$; $X^{\otimes n} Y_T |E\rangle = x_1^{\nu_1} \cdots x_d^{\nu_d} Y_T |E\rangle =: x^\nu Y_T |E\rangle$, where $x^\nu := x_1^{\nu_1} \cdots x_d^{\nu_d}$. Hence, the character value $\operatorname{tr} \mathbf{q}_\lambda(X)$ is the sum of these eigenvalues:

$$\operatorname{tr} \mathbf{q}_\lambda(X) = \sum_\nu K_{\lambda\nu} x^\nu =: s_\lambda(x). \tag{5}$$

Here $K_{\lambda\nu}$ is called the Kostka number and denotes the number of SSYT with weight ν and shape λ. One can show that $K_{\lambda\nu} > 0$ if and only if $\nu \prec \lambda$. We also define here the *Schur polynomial* $s_\lambda(x)$, which is a homogeneous polynomial in d variables of degree $\sum_i \nu_i = n$. Because the character $\operatorname{tr} \mathbf{q}_\lambda(X)$ depends only on the eigenvalues, we will overload notation and denote this character also by $s_\lambda(X)$. For the same reason, it follows that $s_\lambda(XY) = s_\lambda(YX)$. The number of terms of the Schur polynomial is equal to

$$s_\lambda(\mathrm{id}_d) = \operatorname{tr} \mathbf{q}_\lambda(\mathrm{id}_d) = \dim \mathcal{Q}_\lambda = \prod_{i<j} \frac{\lambda_i - \lambda_j + j - i}{j - i}.$$

IV. BOUND ON SCHUR POLYNOMIALS

Lemma 2. *Let ρ and σ be $d \times d$ density matrices. Suppose ρ has rank r. Then, the character function s_λ of the unitary group representation labeled by Young diagram λ satisfies*

$$s_\lambda(\rho\sigma) \begin{cases} \le (\dim \mathcal{Q}_\lambda) e^{-2nH(\bar\lambda)} F^{2n} \\ = 0 \quad \textit{if} \quad \lambda_{r+1} > 0, \end{cases} \tag{6}$$

where

$$F = F(\rho, \sigma) = \operatorname{tr} \sqrt{\sqrt{\rho}\, \sigma \sqrt{\rho}} \tag{7}$$

is the fidelity, and $H(\bar\lambda) = -\sum_i \bar\lambda_i \ln \bar\lambda_i$ is the Shannon entropy of $\bar\lambda = \lambda/n$.

Proof. Consider a positive semi-definite matrix X and a number $k \ge 0$. The largest term in the Schur polynomial $s_\lambda(X^k)$ at eigenvalues $x_1 \ge \cdots \ge x_d \ge 0$ of X is

$$x_1^{k\lambda_1} \cdots x_d^{k\lambda_d} = e^{-nkH(\bar\lambda)} e^{-nkD(\bar\lambda \| \bar x)} (\operatorname{tr} X)^{kn}$$

where $\bar x = (x_1, \ldots, x_d)/\operatorname{tr}(X)$, and $D(p\|q) = \sum_i p_i \ln(p_i/q_i)$ is the relative entropy. This is because majorization implies that

$$\max_{\nu \prec \lambda} x^\nu = x^\lambda,$$

i.e. the maximum is attained by putting the largest number x_1 with the largest possible exponent $\nu_1 = \lambda_1$ and the second largest x_2 with $\nu_2 = \lambda_2$ and so on, subject to the majorization condition $\nu \prec \lambda$.

It follows that

$$s_\lambda(X^k) \le \dim \mathcal{Q}_\lambda \cdot e^{-nkH(\bar\lambda)} e^{-nkD(\bar\lambda\|\bar x)} (\operatorname{tr} X)^{kn}. \tag{8}$$

Now, we set $X = \sqrt{\sqrt{\rho}\, \sigma \sqrt{\rho}}$ and observe $s_\lambda(\rho\sigma) = s_\lambda(X^2)$. Using the fact that $D(\bar\lambda\|\bar x)$ is always non-negative and $= +\infty$ when the rank of $\bar\lambda$ is larger than that of $\bar x$, we arrive at Eq. (6) □

Note that since $s_\lambda(\bar\lambda)$ is a sum of non-negative terms, it is lower bounded by its largest term:

$$s_\lambda(\bar\lambda) \ge e^{-nH(\bar\lambda)}. \tag{9}$$

V. TOMOGRAPHY

Suppose we are given with $\rho^{\otimes n}$, n copies of an unknown density matrix ρ. What is the best strategy to learn about ρ? The input state has a trivial symmetry $\mathbb{S}_n$ under the permutations of the tensor factors. So, the POVM elements of the optimal strategy can be taken to commute with P_π without loss of generality. Additionally since we do not assume any distribution over ρ, our measurement should not perform differently when ρ is replaced by $U\rho U^\dagger$. This means that if M_σ is the outcome corresponding to σ then we should have

$$M_{U\sigma U^\dagger} = (U^\dagger)^{\otimes n} M_\sigma U^{\otimes n}.$$

These observations, along with the Schur-Weyl decomposition, motivate us to define positive semi-definite operators

$$M(\lambda, U) := \frac{\dim \mathcal{Q}_\lambda}{s_\lambda(\bar{\lambda})} \Pi_\lambda (U \bar{\lambda} U^\dagger)^{\otimes n} \Pi_\lambda, \tag{10}$$

for each unitary U and Young diagram λ that partitions n with at most d rows. As before, $\bar{\lambda}$ denotes the diagonal matrix with entries λ/n.

We first show that the $M(\lambda, U)\mathrm{d}U$ constitute a POVM, where $\mathrm{d}U$ is the Haar probability measure on $\mathbb{U}(d)$. It suffices to check $\int \mathrm{d}U M(\lambda, U) = \Pi_\lambda$, for $\sum_\lambda \Pi_\lambda = I$. Since $\int \mathrm{d}U M(\lambda, U)$ is invariant under any unitary conjugation or permutation, we only need to check the traces of both sides.

$$\begin{aligned} \int \mathrm{d}U \operatorname{tr} M(\lambda, U) &= \frac{\dim \mathcal{Q}_\lambda \dim \mathcal{P}_\lambda}{s_\lambda(\bar{\lambda})} \int \mathrm{d}U \operatorname{tr} \mathbf{q}_\lambda (U \bar{\lambda} U^\dagger) \\ &= \frac{\dim \mathcal{Q}_\lambda \dim \mathcal{P}_\lambda}{s_\lambda(\bar{\lambda})} \int \mathrm{d}U \operatorname{tr} \mathbf{q}_\lambda (\bar{\lambda}) \\ &= \operatorname{tr} \Pi_\lambda \end{aligned}$$

Next, we bound the probability density of measuring $M(\lambda, U)$. Let $F = F(\rho, U\bar{\lambda}U^\dagger)$ be the fidelity. We claim

$$\operatorname{tr}(M(\lambda, U)\rho^{\otimes n}) \leq (n+1)^{2dr} F^{2n}, \tag{11}$$

where r is the rank of ρ.

To show this, we need a bound on $\dim \mathcal{P}_\lambda$:

$$\dim \mathcal{P}_\lambda \leq e^{nH(\bar{\lambda})}, \tag{12}$$

which has implicitly appeared in [28]. This follows from

$$\dim \mathcal{P}_\lambda \prod_i \bar{\lambda}_i^{\lambda_i} \leq \frac{n!}{\prod_i \lambda_i!} \prod_i \bar{\lambda}_i^{\lambda_i} = \frac{n!}{n^n} \frac{\prod_i \lambda_i^{\lambda_i}}{\prod_i \lambda_i!} \leq 1. \tag{13}$$

The first inequality is by the "hook length formula" [41]. For the last inequality we note that the function $f(z) = z \ln z - \ln \Gamma(z+1)$ satisfies $f(0) = 0$ and $f''(z) > 0$ for $z > 0$ [42]. Hence, $\sum_{i=1}^d f(\lambda_i)$ with $\sum_{i=1}^d \lambda_i = n$ is maximum if and only if $\lambda_1 = n$, in which case the inequality is saturated.

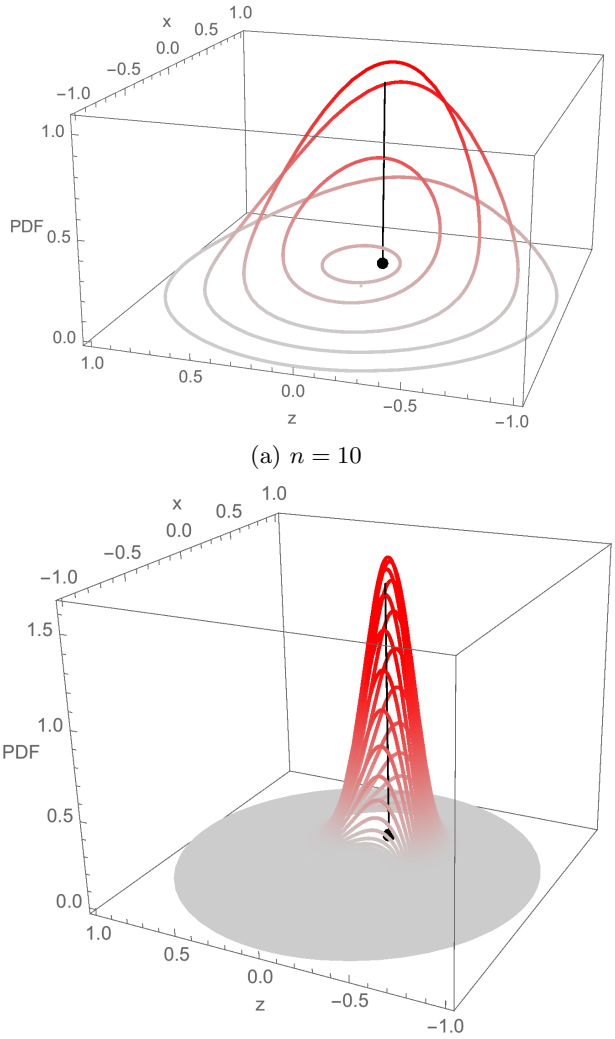
Eqs. (9) and (12) now imply that

$$\begin{aligned} \operatorname{tr}(M(\lambda, U)\rho^{\otimes n}) &= \frac{\dim \mathcal{Q}_\lambda \dim \mathcal{P}_\lambda}{s_\lambda(\bar{\lambda})} s_\lambda(\rho U \bar{\lambda} U^\dagger) \\ &\leq \dim \mathcal{Q}_\lambda \cdot e^{2nH(\bar{\lambda})} s_\lambda(\rho U \bar{\lambda} U^\dagger). \end{aligned}$$

By Eq. (6), this is nonzero only if $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_d = 0$. In this case, we have $\dim \mathcal{Q}_\lambda \leq (n+1)^{dr}$, and arrive at Eq. (11).

The output of our POVM is $\hat{\rho} = U\bar{\lambda}U^\dagger$. The probability of obtaining $\hat{\rho}$ where $\hat{\rho}$ has small fidelity, say infidelity δ, to the true state ρ can be estimated by integrating Eq. (11) over all pairs (λ, U) such that $F(\rho, U\bar{\lambda}U^\dagger) \leq 1 - \delta$. Since $\sum_\lambda \int \mathrm{d}U < (n+1)^d$, we see that

$$\Pr[\ F(\hat{\rho}, \rho) \leq 1 - \delta\ ] \leq (n+1)^{3dr} e^{-2n\delta}. \tag{14}$$



(a) $n = 10$

(b) $n = 100$

FIG. 1. Measurement outcome probability density functions (PDF) of the POVM in Eq. (10) on n copies of a qubit state $\rho = 0.7\,|0\rangle\langle 0| + 0.3\,|1\rangle\langle 1|$, represented by the black delta function. The PDF is plotted over the Bloch states $\hat{\rho} = \frac{1}{2}(I + z\sigma_z + x\sigma_x)$ with $x^2 + z^2 \leq 1$, and is zero except on the circles because output states are of form $U\bar{\lambda}U^\dagger$ where $\bar{\lambda}$ is a density matrix from a discrete set, which becomes finer as n increases. Red is the confidence region, which becomes small for large n.

A. Pretty Good Measurement

Here we propose another POVM that achieves the same (up to constants) sample-complexity for tomography.

Recall that given an ensemble $\{(p_1, \phi_1), \ldots, (p_m, \phi_m)\}$, the PGM has measurement operators $M_i := \bar{\phi}^{-1/2} p_i \phi_i \bar{\phi}^{-1/2}$ with $\bar{\phi} := \sum_i p_i \phi_i$ [35]. A relevant ensemble for us is the one in which ϕ_i is equal to $\sigma_i^{\otimes n}$, and the index i should run over all state space; our ensemble is determined by n and a probability measure $\mathrm{d}\sigma$ on the whole state space $\{\sigma\}$. Demanding the unitary

invariance of $\mathrm{d}\sigma$, we have

$$\bar{\phi} = \int \mathrm{d}\sigma\ \sigma^{\otimes n} = \sum_\lambda \frac{\int \mathrm{d}\sigma s_\lambda(\sigma)}{\dim \mathcal{Q}_\lambda}\Pi_\lambda,$$
$$M_\sigma \mathrm{d}\sigma = \sum_\lambda \frac{\dim \mathcal{Q}_\lambda}{\mathbb{E}\, s_\lambda}\Pi_\lambda \sigma^{\otimes n}\Pi_\lambda \mathrm{d}\sigma, \tag{15}$$

where $\mathbb{E}\, s_\lambda = \int \mathrm{d}\sigma s_\lambda(\sigma)$. It follows that the probability density of measuring M_σ given a state ρ of rank at most r is

$$\begin{aligned}\mathrm{tr}(M_\sigma \rho^{\otimes n})\mathrm{d}\sigma &= \sum_\lambda \frac{(\dim \mathcal{Q}_\lambda \cdot \dim \mathcal{P}_\lambda) s_\lambda(\sigma\rho)}{\mathbb{E}\, s_\lambda}\mathrm{d}\sigma \\ &\le \sum_{\lambda:\lambda_{r+1}=0} \frac{(\dim \mathcal{Q}_\lambda)^2}{e^{nH(\bar{\lambda})}\,\mathbb{E}\, s_\lambda} F^{2n}\mathrm{d}\sigma\end{aligned}$$

where the inequality is by Eq. (6) and (12). This is the same scaling in n up to constants as Eq. (11), provided

$$e^{nH(\bar{\lambda})}\,\mathbb{E}\, s_\lambda \ge (nd)^{-O(dr)}.$$

Indeed we show that this is the case if we choose a uniform distribution over the simplex of spectra of σ. First, we bound the Schur polynomial by its largest term:

$$\int \mathrm{d}\sigma s_\lambda(\sigma) \ge \frac{1}{f_d(\vec{\lambda}=0)} \underbrace{\int_{s_i\ge 0,\ \sum_i s_i = 1} s_1^{\lambda_1}\cdots s_d^{\lambda_d}\mathrm{d}s}_{f_d(\vec{\lambda})}.$$

By writing the integral explicitly, we see that

$$f_d(\lambda_1,\ldots,\lambda_d) = f_2\left(\lambda_1, d-2+\sum_{i=2}^d \lambda_i\right) f_{d-1}(\lambda_2,\ldots,\lambda_d),$$
$$f_2(a,b) = \frac{a!b!}{(a+b+1)!} \qquad \text{(Eq. (16) below)}.$$

This implies that $f_d(\vec{\lambda}) = \lambda_1!\cdots\lambda_d!/(n+d-1)!$. (We just calculated the normalization factor for the Dirichlet distribution.) Hence,

$$\begin{aligned}e^{nH(\bar{\lambda})}\int \mathrm{d}\sigma\ s_\lambda(\sigma) &\ge e^{nH(\bar{\lambda})}\frac{\lambda_1!\cdots\lambda_d!(d-1)!}{(n+d-1)!} \\ &\ge (n+d)^{-d},\end{aligned}$$

where in the second inequality we use Eq. (13). We conclude that this PGM defined by the uniform spectrum distribution achieves the same bound (up to constants) on the sufficient number of copies for tomography.

Both POVMs in Eqs. (10) and (15) are inspired by the pretty good measurement, and indeed the measurement operator corresponding to the estimate σ is like a distorted version of $\sigma^{\otimes n}$. Variants of the PGM have been proposed in which the measurement operators are distorted versions of higher powers of the state $p_i\sigma_i$, i.e. $M_i = X^{-1/2}(p_i\sigma_i)^k X^{-1/2}$ where $X \equiv \sum_i (p_i\sigma_i)^k$. When $k=1$ this is the PGM, but the cases $k=2$ and $k=3$ have also been found useful in specific settings; see [43] for a review. If we take $k\to\infty$ here then this corresponds precisely to the Keyl "rotated-highest-weight" strategy [16]. It is possible that this framework could be used to formally compare the performance of these different strategies.

A definite integral. Here we show for $a,b>0$

$$\int_0^1 u^{a-1}(1-u)^{b-1}\mathrm{d}u = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{16}$$

Let $x = ut \ge 0$ and $y = (1-u)t \ge 0$. The Jacobian is $|\partial(x,y)/\partial(u,t)| = t$. Then,

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty\int_0^\infty \mathrm{d}x\mathrm{d}y\ x^{a-1}y^{b-1}e^{-x-y} \\ &= \int_0^1 \mathrm{d}u \int_0^\infty \mathrm{d}t\ t^{a+b-1}u^{a-1}(1-u)^{b-1}e^{-t} \\ &= \Gamma(a+b)\int_0^1 \mathrm{d}u\ u^{a-1}(1-u)^{b-1}.\end{aligned}$$

□

VI. LOWER BOUNDS

Theorem 3. *Let $\epsilon\in(0,1)$ and $\eta\in(0,1)$. Suppose there exists a POVM $\{M_\sigma \mathrm{d}\sigma\}$ on $(\mathbb{C}^d)^{\otimes n}$ such that for any state $\rho\in\mathbb{C}^{d\times d}$ with rank $\le r$,*

$$\int_{\frac{1}{2}\|\sigma-\rho\|_1\le\epsilon/2} \mathrm{d}\sigma\,\mathrm{tr}[M_\sigma\rho^{\otimes n}] \ge 1-\eta. \tag{17}$$

Then,

$$n \ge C\frac{dr}{\epsilon^2}\frac{(1-\epsilon)^2}{\ln(d/r\epsilon)}$$

for C a constant depending only on η. In addition, if $r=d$, then

$$n \ge C\frac{d^2}{\epsilon^2}(1-\epsilon)^2$$

for C a constant depending only on η.

This theorem implies that achieving infidelity $\delta = 1-F$ requires $n \ge \tilde{\Omega}(dr/\delta)$. For both trace distance and fidelity these lower bounds match our upper bounds up to the log factors.

Let us say that a POVM M_σ on $(\mathbb{C}^d)^{\otimes n}$ is an independent measurement if it is equal to the tensor product of n POVM's $M^{(a)}$ on $\mathbb{C}^d$. Then, we have

Theorem 4. *Let $\delta\in(0,1)$ and $\eta\in(0,1)$. Suppose there exists an independent measurement $M_\sigma\mathrm{d}\sigma$ on $(\mathbb{C}^d)^{\otimes n}$ such that for any state $\rho\in\mathbb{C}^{d\times d}$ with rank $\le r$,*

$$\int_{1-F(\sigma,\rho)\le\delta/4} \mathrm{d}\sigma\,\mathrm{tr}[M_\sigma\rho^{\otimes n}] \ge 1-\eta. \tag{18}$$

Then,

$$n \geq C \frac{dr^2}{\delta^2 \ln(2/\delta)} (1-\delta)^4$$

for C a constant depending only on η. In addition, given $\epsilon \in (0,1)$, if the independent measurement $M_\sigma \mathrm{d}\sigma$ satisfies

$$\int_{\frac{1}{2}\|\rho-\sigma\|_1 \leq \epsilon/2} \mathrm{d}\sigma \, \mathrm{tr}[M_\sigma \rho^{\otimes n}] \geq 1-\eta \tag{19}$$

for any state $\rho \in \mathbb{C}^{d\times d}$ of possibly full rank, then

$$n \geq C \frac{d^3}{\epsilon^2} (1-\epsilon)^2$$

for C a constant depending only on η.

Note that the fidelity lower bound implies trace distance bound

$$n \geq C \frac{dr^2}{\epsilon^2 \ln(2/\epsilon)}.$$

Proof. We will show that any measurement satisfying (17), (18), or (19) will imply the existence of a communication protocol that can reliably send a large message. Holevo's theorem [44] can then be used to obtain a lower bound on n. The independent measurement case is very similar and will be explained at the end of this proof.

Following convention, call the sender Alice and the receiver Bob. We will show in Lemma 5 below that there exists a states $\rho_1, \ldots, \rho_N$ each with rank $\leq r$ such that

$$\frac{1}{2}\|\rho_i - \rho_j\|_1 > \epsilon \quad \forall i \neq j. \tag{20}$$

The set $\{\rho_1, \ldots, \rho_N\}$ is known as an ϵ-packing net. Fix such a net, along with a measurement $\{M_\sigma \mathrm{d}\sigma\}$ satisfying (17).

We will now construct a communication protocol. Alice will choose a message $x \in [N] := \{1, \ldots, N\}$ which she will encode by sending $\rho_x^{\otimes n}$. Bob will use the state estimation scheme $\{M_\sigma\}$ to attempt to guess x. If σ is within $\epsilon/2$ trace distance of some ρ_y then Bob will guess y. By (20), there is always at most one ρ_y satisfying this condition. If no such ρ_y exists, Bob will output failure. This results in the POVM with measurement outcomes

$$\tilde{M}_y = \int_{\frac{1}{2}\|\sigma-\rho_y\|_1 \leq \epsilon/2} \mathrm{d}\sigma M_\sigma \tag{21}$$

$$\tilde{M}_{\mathrm{fail}} = \mathrm{id} - \sum_{y\in[N]} \tilde{M}_y. \tag{22}$$

Define $\Pr[y|x] = \mathrm{tr}[\tilde{M}_y \rho_x^{\otimes n}]$. From (17) we have that $\Pr[x|x] \geq 1-\eta$. In other words, Bob has a $\geq 1-\eta$ chance of correctly decoding Alice's message. By Fano's inequality [45], this implies that

$$I(X:Y) \geq (1-\eta)\ln(N) - \ln(2). \tag{23}$$

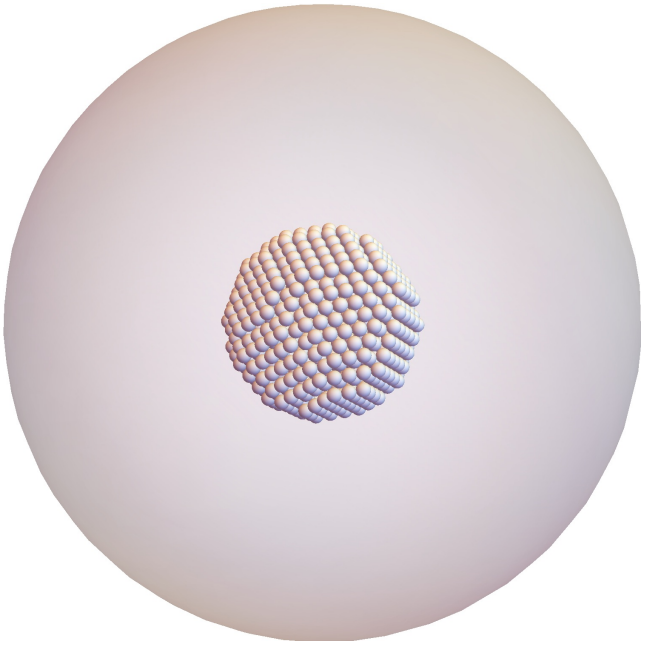


FIG. 2. Packing net in a small region of state space. The centers of the small balls represent states ρ_x that are separated from each other by distance a, which is larger than the resolution of tomography. This enables a communication channel using ρ_x. The packing net has diameter $10a \ll 1$, and contains $N = 10^D$ balls, where $D = d^2 - 1$ is the dimension of the state space, making the channel capacity of order D. Meanwhile, Holevo information for the packing net is proportional to a^2. This establishes our lower bound on the general tomography. For visualization purpose, we depict $D = 3$ case with the packing net centered around the maximally mixed qubit state.

On the other hand, Holevo theorem [44] states that $I(X:Y) \leq \chi$ where χ is the Holevo information:

$$\chi = S\left(\frac{1}{N}\sum_{x\in[N]} \rho_x^{\otimes n}\right) - \frac{1}{N}\sum_{x\in[N]} S(\rho_x^{\otimes n}). \tag{24}$$

In Lemma 5 below we will argue that there exists a packing net with large N and small χ. Specifically, we will bound $\chi \leq n\chi_0$ where

$$\chi_0 = S\left(\mathbb{E}_U U\rho_x U^\dagger\right) - S(\rho_x),$$

for an appropriate Haar random unitary U, and prove $\chi_0 = \tilde{O}(\epsilon^2)$. This will imply that

$$n \geq \frac{(1-\eta)\ln(N) - \ln(2)}{\chi_0}.$$

Our result then follows from Lemma 5 below.

For the independent measurements, Bob has to infer the state based on the measurement outcome distribution from each copy. Hence, the Holevo information must be calculated with respect to the outcome distribution. Since the construction of the states, and the calculation of Holevo information are somewhat similar to those for

the joint measurements, we present a complete proof in Sec. VI C after the proof of Lemma 5. □

Lemma 5. *There exist ϵ-packing nets I,II,III of d-dimensional states (i.e. satisfying (20)) characterized in the following table.*

	rank	$\chi_0/c \leq$	$c \ln N \geq$	restriction
I	r	$\epsilon^2 \ln(d/r\epsilon)$	rd	$\epsilon \leq 2^{-4}$, $r < d/3$
II	d	ϵ^2	d^2	$\epsilon \leq 2^{-3}$, d *even*
III	r	$\ln(d/r)$	$rd(1-\epsilon)$	$r < d(1-\epsilon)/6$

where $c > 0$ is a sufficiently large constant; $c = 1000$ is good enough.

We remark that packing nets of size $\exp(\Omega(dr))$ for rank-r states have been achieved as early as 1981 [46, 47]; see also [48, 49] which used them for applications in communication complexity. These imply an $\Omega(dr)$ lower bound on the number of copies needed when ϵ is constant [39, 48, 49] and has been used in [20] to argue an $\tilde{\Omega}(r^2d^2)$ lower bound on the number of copies needed for constant accuracy using adaptive Pauli measurements. Our main new contribution here is to analyze at the same time the Holevo capacity corresponding to these ensembles, in order to obtain bounds with simultaneously optimal scaling with r, d and ϵ.

A. Probabilistic existence argument

We will define a set of states $\rho_U = U\rho_I U^\dagger$ where U is any element of some subgroup $G \subseteq \mathbb{U}(d)$. Suppose

$$\Pr_U[\ \|\rho_U - \rho_I\|_1 \leq \epsilon\ ] \leq \zeta$$

for Haar random $U \in G$. We wish to find a set $\{U_i\}$ of unitaries with cardinality at least $\lceil 1/\zeta \rceil$ such that $\|\rho_{U_i} - \rho_{U_j}\|_1 > \epsilon$ whenever $i \neq j$. This can be done inductively starting with the singleton $\{I\}$. Since Haar measure is left-invariant, $\Pr_U[\ \|\rho_U - \rho_V\|_1 \leq \epsilon\ ] \leq \zeta$ for any unitary $V \in G$. If $m < \lceil 1/\zeta \rceil$ unitaries are chosen, the probability of choosing a unitary U such that ρ_U is ϵ-close to any previously chosen ρ_{U_i} is at most ζm, which is strictly smaller than 1. This proves the existence of one more desired unitary, and we obtain a set of $\lceil 1/\zeta \rceil$ elements. The probability ζ will be repeatedly estimated using the following fact.

Lemma 6 (Lemma III.5 of Ref. [50])**.** *Let P and Q be projectors on $\mathbb{C}^d$ of rank p and q, respectively. Let $U \in \mathbb{U}(d)$ be Haar random. It holds that*

$$\forall z > 0 : \Pr_U\left[\frac{d}{pq}\operatorname{tr} QUPU^\dagger \geq 1 + z\right] \leq \exp[-pqf(z)],$$
$$\forall z \in (0,1) : \Pr_U\left[\frac{d}{pq}\operatorname{tr} QUPU^\dagger \leq 1 - z\right] \leq \exp[-pqf(-z)],$$

where

$$f(z) = z - \ln(1+z) \geq \begin{cases} (1+z)/2 & z \in [5,\infty) \\ (1-\ln 2)\ z^2 & z \in (-1,1] \\ z^2/2 & z \in (-1,0] \end{cases}.$$

Ref. [50] does not explicitly cover the $z > 1$ case for the first inequality, though it implicitly covered in their proof. See Appendix A.

B. Joint Measurement

This section constitutes the proof of Lemma 5.

1. *Packing net I*

Suppose $3r < d$. Let

$$U = \begin{pmatrix} I_r & 0 & 0 \\ 0 & A_{r\times r} & B_{r\times(d-2r)} \\ 0 & C_{(d-2r)\times r} & D_{(d-2r)\times(d-2r)} \end{pmatrix} \tag{25}$$

be a unitary matrix of $\mathbb{U}(d-r)$ with blocks as indicated, embedded into $\mathbb{U}(d)$. For $0 \leq t \leq 1$, define

$$\rho_{t,I} = \begin{pmatrix} (1-t^2)I_r/r & t\sqrt{1-t^2}I_r/r & 0 \\ t\sqrt{1-t^2}I_r/r & t^2 I_r/r & 0 \\ 0 & 0 & 0_{d-2r} \end{pmatrix}, \tag{26}$$

$\rho_{t,U} = U\rho_{t,I}U^\dagger$.

It is a maximally mixed state on an r-dimensional subspace. We claim that the distance between $\rho_{t,U}$ satisfies

$$\|\rho_{t,U} - \rho_{t,I_{d-r}}\|_1 \geq \frac{t\sqrt{1-t^2}}{r}\operatorname{tr} C^\dagger C \tag{27}$$

where C is as in Eq. (25). To prove this, observe that $\|\rho_{t,U} - \rho_{t,I_{d-r}}\|_1 \geq |\operatorname{tr}[(\rho_{t,U} - \rho_{t,I_{d-r}})V]|$ where

$$V = \begin{pmatrix} A & 0 & BF \\ 0 & E & 0 \\ C & 0 & DF \end{pmatrix}$$

and $E \in \mathbb{U}(r)$ and $F \in \mathbb{U}(d-2r)$ are arbitrary. Abbreviate as $\alpha = (1-t^2)/r$, $\beta = t\sqrt{1-t^2}/r$, and $\gamma = t^2/r$. Expanding the formula,

$$\begin{aligned} &\operatorname{tr}[(\rho_{t,U} - \rho_{t,I_{d-r}})V] \\ &= \operatorname{tr}\left[\begin{pmatrix} 0 & \beta(A^\dagger - I) & \beta C^\dagger \\ \beta(A-I) & \gamma(AA^\dagger - I) & \gamma AC^\dagger \\ \beta C & \gamma CA^\dagger & \gamma CC^\dagger \end{pmatrix}\begin{pmatrix} A & 0 & BF \\ 0 & E & 0 \\ C & 0 & DF \end{pmatrix}\right] \\ &= \operatorname{tr}\begin{pmatrix} \beta C^\dagger C & \star & \star \\ \star & (\gamma AA^\dagger - I)E & \star \\ \star & \star & (\beta CB + \gamma CC^\dagger D)F \end{pmatrix}. \end{aligned}$$

For some unitary E and F, the trace of the last two entries become the trace norm of the matrices in the parentheses, which are non-negative. This proves Eq. (27).

Lemma 7. *If $0 < t < 1/2$ and $r < d/3$, there exists a finite subset $\{U_i\} \subset \mathbb{U}(d-r)$ of cardinality $N \geq \exp(dr/54)$ such that $\|\rho_{t,U_i} - \rho_{t,U_j}\|_1 > t/4$ for any $i \neq j$. The Holevo χ_0 of $\{\rho_{t,U_i}\}_{i=1}^N$ fulfills $\chi_0 \leq t^2 \ln \frac{ed}{t^2 r}$.*

Proof. Lemma 6 states that if U is a Haar random unitary matrix of dimension k, then any $k_1 \times k_2$ subblock K of U satisfies

$$\Pr\left[\frac{k}{k_1 k_2}\operatorname{tr}(K^\dagger K) < 1 - z\right] \leq \exp(-k_1 k_2 z^2/2)$$

for $z \in (0,1)$. Eq. (27) says that $\|\rho_{t,I_{d-r}} - \rho_{t,U}\|_1 \leq t/4$ implies $\frac{d-r}{r(d-2r)}\operatorname{tr} C^\dagger C \leq \frac{1}{\sqrt{3}} < 1 - \frac{1}{3}$. Therefore,

$$\Pr[\|\rho_{t,I_{d-r}} - \rho_{t,U}\|_1 \leq t/4] \leq e^{-r(d-2r)/18} < e^{-rd/54},$$

and we resort to the probabilistic existence argument.

Next, we estimate the Holevo information χ. Since U is unitary, we have $S(\rho_{t,U}) = S(\rho_{t,I_{d-r}}) = \ln r$. By the concavity of entropy, the ensemble average may be replaced with $\bar{\rho}_t = \int \mathrm{d}U \rho_{t,U}$, only to increase the entropy. By Schur's lemma, the matrix $\bar{\rho}_t$ is diagonal, and has entropy

$$S(\bar{\rho}_t) = H(t^2) + (1-t^2)\ln r + t^2 \ln(d-r),$$

where $H(t^2) = -t^2\ln(t^2) - (1-t^2)\ln(1-t^2)$ is the binary entropy. Combining, we have $\chi/n \leq H(t^2) + t^2 \ln \frac{d-r}{r}$. Using $H(z) \leq z\ln(e/z)$, we finish the proof. □

2. *Packing nets II & III*

Assume that d is an even number, and fix a projector $Q = \operatorname{diag}(1,\ldots,1,0,\ldots,0)$ of rank $r \leq d/2$. For any $d \times d$ unitary U and $0 \leq t \leq 1$, define

$$\tau_{t,U} = \frac{1+t}{2r} UQU^\dagger + \frac{1-t}{2(d-r)}(I_d - UQU^\dagger). \tag{28}$$

Given an ensemble $\{\tau_{t,U}\}$, the entropy of the ensemble average is certainly at most $\ln d$. The entropy of $\tau_{t,U}$ is equal to $H((1+t)/2) + \frac{1+t}{2}\ln r + \frac{1-t}{2}\ln(d-r)$, where $H(\cdot)$ is the binary entropy. Therefore, the Holeve χ_0 is bounded as

$$\chi_0 \leq \frac{1}{2}\ln\frac{d^2}{r(d-r)} + \frac{t}{2}\ln\frac{d-r}{r} - H\left(\frac{1+t}{2}\right). \tag{29}$$

Next, if A denotes the upper-left $r \times r$ and C the lower-left $(d-r)\times r$ submatrix of U, we have

$$\begin{aligned} \operatorname{tr} AA^\dagger + \operatorname{tr} CC^\dagger &= r \\ \operatorname{tr} BB^\dagger + \operatorname{tr} DD^\dagger &= d-r \\ \operatorname{tr} CC^\dagger + \operatorname{tr} DD^\dagger &= d-r \end{aligned} \tag{30}$$

and

$$\tau_{t,U} - \tau_{t,I_d} = \begin{pmatrix} \alpha AA^\dagger + \beta BB^\dagger - \alpha I_r & \star \\ \star & \alpha CC^\dagger + \beta DD^\dagger - \beta I_{d-r} \end{pmatrix}$$

where $\alpha = (1+t)/2r$ and $\beta = (1-t)/2(d-r)$. Multiplying a unitary $\operatorname{diag}(-I_r, I_{d-r})$ on the right of $\tau_{t,U} - \tau_{t,I_d}$, we see that

$$\begin{aligned} &\|\tau_{t,U} - \tau_{t,I_d}\|_1 \\ &\geq \alpha\operatorname{tr}(CC^\dagger - AA^\dagger) + \beta(DD^\dagger - BB^\dagger) + (d-r)\beta - r\alpha \\ &= 2(\alpha-\beta)\operatorname{tr}(CC^\dagger) \qquad \text{by Eq. (30)} \\ &= \left(\frac{1+t}{r} - \frac{1-t}{d-r}\right)\operatorname{tr} CC^\dagger. \end{aligned} \tag{31}$$

Lemma 8. *Suppose $r = d/2$. Then, there exists a finite subset $\{U_i\} \subset \mathbb{U}(d)$ of cardinality $N \geq \exp(d^2/32)$ such that $\|\tau_{t,U_i} - \tau_{t,U_j}\|_1 > t/2$ for any $i \neq j$. The Holevo χ_0 fulfills $\chi_0 \leq t^2$.*

Proof. Eq. (29) becomes $\chi/n \leq \ln 2 - H((1+t)/2) \leq t^2$. Eq. (31) says that if $\|\tau_{t,U} - \tau_{t,I}\|_1 \leq t/2$, then $(4/d)\operatorname{tr} CC^\dagger \leq 1/2$. Lemma 6 states that this happens with probability at most $\exp(-d^2/32)$. The probabilistic existence argument applies. □

Lemma 9. *Set $t = 1$. Suppose $\epsilon \in (0,1)$, and $r < d(1-\epsilon)/6$. Then, there exists a finite subset $\{U_i\} \subset \mathbb{U}(d)$ of cardinality $N \geq \exp((1-\epsilon)rd/2)$ such that $\|\tau_{1,U_i} - \tau_{1,U_j}\|_1 > 2\epsilon$ for any $i \neq j$. The Holevo χ_0 fulfills $\chi_0 \leq \ln(d/r)$.*

Proof. Eq. (29) becomes $\chi_0 \leq \ln(d/r)$. Eq. (31) says that if $\|\tau_{t,U} - \tau_{t,I}\|_1 \leq 2\epsilon$, then $\frac{d}{r^2}\operatorname{tr} AA^\dagger \geq (1-\epsilon)d/r$, which is greater than 6 when $r < d(1-\epsilon)/6$. By Lemma 6, this happens with probability at most $\exp(-r^2(1-\epsilon)d/2r) = \exp(-rd(1-\epsilon)/2)$. The probabilistic existence argument applies. □

C. Independent Measurement

Proof of Theorem 4 continued from p. 9. Since $\sqrt{1-F}$ is a metric (Bures metric) on the space of states, if there is a set of states ρ_i such that $1 - F(\rho_i, \rho_j) > \delta$ for all $i \neq j$, then for any ρ there is at most one ρ_i such that $1 - F(\rho_i, \rho) \leq \delta/4$. In the regime where δ is close to 1, we can use Packing Net III analyzed in Lemma 9. Since $1 - T \geq 1 - \sqrt{1-F^2} \geq F^2/2$, we obtain a packing net of cardinality $N = \exp(\Omega(rd(1-\delta)^4))$ in which every state has rank at most r and every pair has infidelity at least $\delta \in (0,1)$.

In order to compute Holevo information and to account for the small δ regime, we consider the following set of

states. Define for $t \in (0,1)$ and $U \in \mathbb{U}(d-1)$

$$\omega_{t,I} = \begin{pmatrix} (1-t) & & \\ & tI_r/r & \\ & & 0_{d-r-1} \end{pmatrix} \tag{32}$$

$$\omega_{t,U} = U\omega_{t,I}U^\dagger \tag{33}$$

where U is embedded into $\mathbb{U}(d)$ similarly as in Eq. (25). $\omega_{t,U}$ has rank $r+1 < d$. Applying the defining formula $F = \operatorname{tr}\sqrt{\sqrt{\omega_{t,I}}\omega_{t,U}\sqrt{\omega_{t,I}}}$ with the observation that $\omega_{t,U}$ is a mixture of two orthogonal states, we obtain

$$1 - F(\omega_{t,U}, \omega_{t,I}) = t(1 - F(\tau'_I, \tau'_U)) \tag{34}$$

where $\tau'_U = U\tau'_I U^\dagger$ is the $(d-1)$-dimensional state that is maximally mixed on an r-dimensional subspace. (τ'_U is equal to $\tau_{t=1,U}$ of Eq. (28) except the size.) Since

$$T^2 \le 2(1-F)$$

by Eq. (1), we can apply the probabilistic existence argument to find a set of states of cardinality $\exp(\Omega(rd))$ that are $\delta = \Omega(t)$-separated in infidelity.

For the full rank case where the accuracy is measured in the trace distance, we use Packing Net II analyzed in Lemma 8, from which we know there are $\exp(\Omega(d^2))$ states separated by the trace distance $\Omega(t)$.

Bounds for the Holevo information are supplied by the following two lemmas. □

Lemma 10. *Suppose $\vec{M}^{(a)}$ for each $a = 1, \dots, n$ is a POVM on $\mathbb{C}^d$. Consider $\tau_{t,U}$ in Eq. (28) with $r = d/2$. For any distribution of unitaries $\{U_j\} \subseteq \mathbb{U}(d)$ there exists $W \in \mathbb{U}(d)$ such that the Holevo information of*

$$\left\{ \operatorname{tr}\left((\tau_{t,WU_j})^{\otimes n} \bigotimes_{a=1}^{n} \vec{M}^{(a)} \right) \right\}$$

is at most nt^2/d.

Proof. The first term of the Holevo information χ is the Shannon entropy of the distribution

$$\mathfrak{p} = \mathop{\mathbb{E}}_{\tau} \operatorname{tr}\left(\tau^{\otimes n} \bigotimes_{a=1}^{n} \vec{M}^{(a)} \right)$$

whose marginal is equal to $\mathbb{E}_\tau \operatorname{tr}(\tau \vec{M}^{(a)})$. By the subadditivity of entropy, we have

$$H(\mathfrak{p}) \le \sum_{a=1}^{n} H(\mathop{\mathbb{E}}_{\tau} \operatorname{tr}(\tau \vec{M}^{(a)})).$$

It follows that

$$\chi\{W\tau_j W^\dagger\} \le \sum_{a=1}^{n} \chi_a\{W\tau_j W^\dagger\} \tag{35}$$

where the subscript a means with respect to $\vec{M}^{(a)}$. Minimizing the right-hand side by varying W, we see there exists W such that

$$\chi\{\rho_{WU_j}\} \le \min_V \sum_a \chi_a\{\tau_{VU_j}\}.$$

The minimum on the right-hand side is at most the average over V from the Haar measure.

$$\min_V \sum_a \chi_a\{\tau_{VU_j}\} \le \mathop{\mathbb{E}}_V \sum_a \chi_a\{\tau_{VU_j}\}.$$

By concavity of entropy,

$$\mathop{\mathbb{E}}_V \sum_a \chi_a\{\tau_{VU_j}\} \le \sum_a \chi_a\{\tau_{t,U} : \text{ Haar uniform } U \in \mathbb{U}(d)\}.$$

Hence, it suffices to prove the lemmas when the initial distribution of U_j is Haar uniform, which we assume hereafter. In addition, it suffices to consider rank-1 POVM elements since one can always decompose a POVM element into rank-1 projectors of some positive weight. Let each POVM element be $M_i = w_i d\,|a_i\rangle\langle a_i|$.

The outcome probability is

$$\begin{aligned} p_i \equiv \operatorname{tr}(M_i \tau_{t,U}) &= w_i d \left(\frac{2t}{d} \operatorname{tr}(P_1^{(i)} U P_{d/2} U^\dagger) + \frac{1-t}{d} \right) \\ &=: w_i (1 - t + t Z^{(i)}_{d,d/2}) \end{aligned}$$

where $P_1^{(i)} = |a_i\rangle\langle a_i|$. Since $\mathbb{E}_U Z^{(i)}_{d,d/2} = 1$ for any i (see Eq. (38) below), the Holevo information per copy is

$$\begin{aligned} \chi_a &= \sum_{i=1}^{m} -\mathbb{E}[p_i] \ln \mathbb{E}[p_i] + \mathop{\mathbb{E}}_U [p_i \ln p_i] \\ &= \sum_{i=1}^{m} w_i \mathop{\mathbb{E}}_U \left[(1 - t + t Z^{(i)}_{d,d/2}) \ln(1 - t + t Z^{(i)}_{d,d/2}) \right] \\ &\le \sum_{i=1}^{m} w_i t^2 \left(\mathop{\mathbb{E}}_U \left[(Z^{(i)}_{d,d/2})^2 \right] - 1 \right). \end{aligned}$$

Since $\mathbb{E}_U (Z^{(i)}_{d,d/2})^2 = (1 + 2/d)/(1 + 1/d)$ for any i (see Eq. (39) below), we have

$$\chi(\tau_{t,U}) \le \sum_{a=1}^{n} \chi_a \le nt^2/(d+1). \tag{36}$$

This completes the proof of Lemma 10. □

Random variable Z. Define the random variable $Z_{n,m}$ to be

$$Z_{n,m} := \frac{n}{m} \frac{x_1^2 + \cdots + x_{2m}^2}{x_1^2 + \cdots + x_{2n}^2} \tag{37}$$

where x_i are independent identical Gaussians with mean 0 and variance 1/2. Here we show

$$\mathbb{E}\, Z_{n,m} = 1, \tag{38}$$

$$\mathbb{E}\, Z_{n,m}^2 = \frac{1 + 1/m}{1 + 1/n}, \tag{39}$$

by deriving the probability density function $p(Z_{n,m} = z)$ on $[0, n/m]$

$$p(z) = \frac{m\Gamma(n)}{n\Gamma(n-m)\Gamma(m)} \left(\frac{mz}{n}\right)^{m-1} \left(1 - \frac{mz}{n}\right)^{n-m-1}. \tag{40}$$

To this end, let $x = (x_1, \ldots, x_{2m})$ and $y = (x_{2m+1}, \ldots, x_{2n})$ be Cartesian coordinates for $\mathbb{R}^{2n}$. Let $\mathrm{d}^{2m-1}\Omega_x$ and $\mathrm{d}^{2n-2m-1}\Omega_y$ be the solid angle elements of respective dimensions. Then the volume form $\mathrm{d}V = \mathrm{d}^{2m}x\mathrm{d}^{2n-2m}y$ is equal to $|x|^{2m-1}|y|^{2n-2m-1}\mathrm{d}|x|\mathrm{d}|y|\mathrm{d}\Omega_x\mathrm{d}\Omega_y$. Defining new variables r, θ by $|x| = r\cos\theta$ and $|y| = r\sin\theta$ ($\theta \in [0, \pi/2]$), we see that the $(2n-1)$-dimensional solid angle element is

$$\left.\frac{\mathrm{d}V}{\mathrm{d}r}\right|_{r=1} = \cos^{2m-1}\theta \sin^{2n-2m-1}\theta \mathrm{d}\theta\mathrm{d}\Omega_x\mathrm{d}\Omega_y$$

Since our variable $Z_{n,m} = (n/m)\cos^2\theta =: (n/m)u$ is a function of θ only, we integrate out $\mathrm{d}\Omega_x\mathrm{d}\Omega_y$, and use the relation $\mathrm{d}\theta = u^{-1/2}(1-u)^{-1/2}\mathrm{d}u$ to arrive at Eq. (40) after normalization using Eq. (16). □

Lemma 11. *Let $t \in (0, 1/3)$ and $d \geq 3$. Suppose $\vec{M}^{(a)}$ for each $a = 1, \ldots, n$ is a POVM on $\mathbb{C}^d$. For any distribution of unitaries $\{U_j\} \subseteq \mathbb{U}(d-1)$ there exists $W \in \mathbb{U}(d)$ such that the Holevo information of*

$$\left\{\mathrm{tr}\left((W\omega_{t,U_j}W^\dagger)^{\otimes n} \bigotimes_{a=1}^{n} \vec{M}^{(a)}\right)\right\}$$

is at most $4(nt^2/r)\ln(2/t)$ where ω is as in Eq. (33).

Proof. The first stage of the proof is similar to that of Lemma 10; we use the freedom W and consider $W\omega_{t,VU_j}W^\dagger$ for some $W \in \mathbb{U}(d)$ and $V \in \mathbb{U}(d-1)$. By varying V, we may assume that our ensemble $\mathcal{M}_W$ is

$$\mathcal{M}_W = \{W\omega_{t,U}W^\dagger : U \in \mathbb{U}(d-1) \text{ is Haar random. }\}$$

and we will estimate the Holevo information per copy χ_a of this ensemble.

There is still remaining freedom to choose $\mathcal{M}_W$ using $W \in \mathbb{U}(d)$. Certainly,

$$\min_W \chi_a(\mathcal{M}_W) \leq \mathop{\mathbb{E}}_W \chi_a(\mathcal{M}_W)$$

where the average over W is with respect to Haar random W and the inequality is saturated when our POVM consists of rank-1 projectors $|v\rangle\langle v|$ from Haar uniform distribution, which we assume hereafter.

The outcome probability density is

$$\begin{aligned} p(|v\rangle, U) &\equiv \mathrm{tr}(d\,|v\rangle\langle v|\,\omega_{t,U}) \\ &= d(1-t)|v_1|^2 + \frac{td}{r}(1-|v_1|^2)\,\mathrm{tr}(UP_rU^\dagger P_1) \\ &= (1-t)\underbrace{Z_{d,1}}_{v} + t\frac{d-Z_{d,1}}{d-1}\underbrace{Z'_{d-1,r}}_{z} \end{aligned}$$

where v_1 is one component of the vector v, P_r and P_1 are r- and 1-dimensional projectors, respectively, and in the third line we used the notation in Eq. (37). We use the notation $Z_{d,1}, Z'_{d-1,r}$ to mean two independent random variables defined according to (37) for appropriate choices of n, m. Since we do not use all the degrees of freedom in U, we can think of v, z as our random variables (distributed according to $Z_{d,1}, Z'_{d-1,r}$ respectively), corresponding to outcome probability

$$p(v,z) = (1-t)v + t\frac{d-v}{d-1}z. \tag{41}$$

Now,

$$\chi_a = \mathop{\mathbb{E}}_v\left[-p(v, \mathop{\mathbb{E}}_z[z])\ln p(v, \mathop{\mathbb{E}}_z[z]) + \mathop{\mathbb{E}}_z[p(v,z)\ln p(v,z)]\right],$$

and we use $\mathbb{E}_z\, z = 1, \mathbb{E}_z\, z^2 = (1+1/r)/(1+1/(d-1))$, and

$$\ln p(v,z) \leq \ln p(v,1) + \frac{p(v,z)-p(v,1)}{p(v,1)},$$

to obtain

$$\begin{aligned} \chi_a &\leq \frac{t^2(d-r-1)}{rd(d-1)} \cdot \mathop{\mathbb{E}}_v \frac{(d-v)^2}{td + (d-td-1)v} \\ &\leq \frac{t^2}{r} \cdot \mathop{\mathbb{E}}_v \frac{1}{t + v/3} \end{aligned}$$

where the last line is because $t < 1/3$ and $d \geq 3$.

The probability density function of $v = Z_{d,1}$ is given by Eq. (40)

$$f(v) = (1-1/d)(1-v/d)^{d-2} < 1.$$

Since $t > 0$,

$$\begin{aligned} \mathop{\mathbb{E}}_v \frac{1}{t+v/3} &= \left(\int_0^3 + \int_3^d\right)\frac{f(v)}{t+v/3}\mathrm{d}v \\ &\leq \int_0^3 \frac{\mathrm{d}v}{t+v/3} + \int_3^d f(v)\mathrm{d}v \\ &< 3\ln(1+1/t) + 1 \\ &< 4\ln(2/t). \end{aligned}$$

This completes the proof. □

VII. IMPLEMENTATION ON A QUANTUM COMPUTER

In this section we informally describe how our tomography strategy can be implemented in time $n^{O(dr)}$ on a quantum computer.

Our measurement involves a POVM with a continuously infinite number of outcomes. However, it can be approximated with a finite POVM using ideas from [51]. The first step is to measure λ, as proposed by Keyl-Werner [26].

This can be done efficiently using the Schur transform [52] or the quantum Fourier transform over the symmetric group [53, 54].

Next, we would like to find a collection of unitaries $U_1, \ldots, U_m$ such that

$$\frac{1}{m}\sum_{i=1}^{m} M(\lambda, U_i) \approx \Pi_\lambda.$$

This can be done by choosing $m = \tilde{O}(\dim \mathcal{Q}_\lambda/\epsilon^2)$ random unitaries, as proven in [51], which in turn was based on [38]). The resulting measurement can be implemented by the isometry

$$V = m^{-1/2}\sum_{i=1}^{m} \sqrt{M(\lambda, U_i)} \otimes |i\rangle\,.$$

Using the Schur transform, this reduces to performing the isometry

$$\tilde{V} = C\sum_{i=1}^{m} \sqrt{\mathbf{q}_\lambda(U_i \bar{\lambda} U_i^\dagger)} \otimes |i\rangle\,,$$

where C is a normalizing constant. This isometry can be implemented using $O((\dim \mathcal{Q}_\lambda)^2 m^2)$ gates [55], which is $\tilde{O}(n^{2dr}/\epsilon^2)$.

We conjecture that run-time $\mathrm{poly}(n, d, \ln(1/\epsilon))$ is possible, but do not know how to achieve this, even in the relatively simple case of $r = 1$.

VIII. DISCUSSION

The sample complexity of the general quantum tomography problem is nearly resolved here. It is confirmed up to logarithmic factors that one only needs as many copies as the number of unknown parameters if one can perform joint measurements. In addition, we have shown information-theoretically that this optimal measurement *cannot* be a combination of independent measurements. Our result raises an important question on the performance of *adaptive* measurements, where an individual copy is measured at a time, but each measurement may utilize the history of outcomes on other copies. Is there an asymptotic separation between the power of adaptive and collective measurements? Another open problem is whether our joint measurement scheme can be implemented efficiently on a quantum computer; we briefly remark that the implementation is possible in a polynomial time in n for a fixed d, but dependence on d is exponential. There is a method to extract the eigenvectors of a small-rank density matrix on a quantum computer efficiently [56], but it remains challenging to convert the eigenvector into a classical description.

An independent and concurrent work [39] analyzes Keyl's measurement strategy [16], and proves that it only requires $n = O(dr/\epsilon^2)$ copies to achieve ϵ accuracy in trace distance. This improves on our corollary for trace distance by removing the logarithmic factor, but does not imply our fidelity bound, which is incomparable to theirs.

ACKNOWLEDGMENTS

We thank Robin Blume-Kohout, Steve Flammia, Masahito Hayashi, Debbie Leung, and John Watrous for discussions. We also thank Ryan O'Donnell and John Wright for sharing their draft of [39] with us. JH is supported by the Pappalardo Fellowship in Physics while at MIT. AWH was funded by NSF grants CCF-1111382 and CCF-1452616 and ARO contract W911NF-12-1-0486. ZJ and NY's research was supported by NSERC, NSERC DAS, CRC, and CIFAR. XW's research was funded by ARO contract W911NF-12-1-0486 and by the NSF Waterman Award of Scott Aaronson. Part of the research was conducted when XW was visiting Institute for Quantum Computing (IQC), University of Waterloo and XW thanks IQC for its hospitality.

Appendix A: Overlap of random projectors

Here, we provide a self-contained proof of Lemma 6 (Lemma III.5 of Ref. [50]). We follow the ideas of Ref. [50] and [57].

Lemma 12. *Let $\mathcal{D}$ be the set of all $d \times d$ normalized density matrices of rank p, and Δ be the set of all probability vectors η of length p. Suppose $\mathcal{D}$ has a $\mathbb{U}(d)$-invariant probability measure $\mathrm{d}\rho$. Then, there exists a permutation-symmetric probability measure $\mathrm{d}\eta$ on Δ such that*

$$\iint \mathrm{d}\eta\ \mathrm{d}U\ f(U\eta U^\dagger) = \int \mathrm{d}\rho\ f(\rho)$$

for any continuous function f on $\mathcal{D}$ where $\mathrm{d}U$ is the normalized Haar measure on $\mathbb{U}(d)$, and η in between U and $U^\dagger$ denotes the diagonal matrix with entries $(\eta_1, \ldots, \eta_p, 0, \ldots, 0)$.

This means that the eigenvalues and the eigenvectors can be treated as if they were "independent random variables." Strictly speaking, $\mathrm{d}\eta$ and $\mathrm{d}U$ are *not* derived from ρ; we just find that they induce the measure $\mathrm{d}\rho$ on $\mathcal{D}$ by the map $(\eta, U) \mapsto U\eta U^\dagger$.

Proof. Since *sorted* eigenvalues are continuous functions of the matrix, we have a map $\lambda : \mathcal{D} \to \Delta^\downarrow$, which induces a measure $\mathrm{d}\lambda$ on $\Delta^\downarrow$, the set of all sorted non-negative p real numbers summing to 1. The defining equation for the induced measure is $\int \mathrm{d}\rho\ g(\lambda(\rho)) = \int \mathrm{d}\lambda\ g(\lambda)$ for any continuous function g. Here, we have identified a vector with a diagonal matrix padded with $(d-p)$ zeros. Define

$$\bar{f}(\rho) = \int \mathrm{d}U f(U\rho U^\dagger)$$

so that $\bar{f}(\rho) = \bar{f}(V\rho V^\dagger)$ for any $V \in \mathbb{U}(d)$. Since $\mathrm{d}\rho$ is unitary invariant, $\int \mathrm{d}\rho f(\rho) = \int \mathrm{d}\rho f(U\rho U^\dagger)$. Integrating the both sides over U, $\int \mathrm{d}\rho f(\rho) = \int \mathrm{d}U \int \mathrm{d}\rho f(U\rho U^\dagger) = \int \mathrm{d}\rho \bar{f}(\rho)$. (All spaces are compact, so integration order

never matters.) We can now prove an analogous version of the lemma for $\Delta^{\downarrow}$:

$$\int \mathrm{d}\rho f(\rho) = \int \mathrm{d}\rho \bar{f}(\rho) = \int \mathrm{d}\rho \bar{f}(\lambda(\rho)) = \int \mathrm{d}\lambda \bar{f}(\lambda) = \iint \mathrm{d}\lambda \mathrm{d}U f(U\lambda U^\dagger)$$

In order to finish the proof, all we need is to divide Δ into $p!$ pieces, each of which is mapped to $\Delta^{\downarrow}$ by permuting components up to measure zero sets, and assign measure to each piece by $\mathrm{d}\lambda/p!$. Thus defined $\mathrm{d}\eta$ on Δ is permutation-invariant. □

Lemma 13. *Let $x_1, x_2, \ldots$ be independent gaussian random variables with mean 0 and variance $\frac{1}{2}$. Let U be a Haar random unitary of dimension d, and P and Q be d-dimensional projectors of rank p and q, respectively. For any real number ξ, it holds that*

$$\mathbb{E}_{x_i} \exp\left[\xi \sum_{i=1}^{2pq} x_i^2\right] \geq \mathbb{E}_U \exp\left[\xi d \operatorname{tr}(QUPU^\dagger)\right].$$

Proof. Consider $\mathbb{C}^{dp} = \mathbb{C}^d \otimes \mathbb{C}^p$, and define $Q' = Q \otimes I_p$ to be the projector of rank qp. Without loss of generality, we assume that P, Q are diagonal. The random tuple $(x_1, \ldots, x_{2dp})$ has the probability density $\frac{1}{\pi^{dp}} \exp(-r^2) \mathrm{d}^{2dp}x$ where $r^2 = \sum_{i=1}^{2dp} x_i^2$. This means in particular that the magnitude variable r and the direction variable $\hat{x} = (x_1, \ldots, x_{2dp})/r$ are independent. The direction variable $\hat{x}$ defines a normalized pure state $|\hat{x}\rangle$ on $\mathbb{C}^d \otimes \mathbb{C}^p$, and the sum $\sum_{i=1}^{2pq} \hat{x}_i^2$ can be regarded as the squared norm of $Q' |\hat{x}\rangle$.

$$\sum_{i=1}^{2pq} x_i^2 = r^2 \langle \hat{x}| Q' |\hat{x}\rangle = r^2 \operatorname{tr} Q\rho$$

where ρ is the reduced density matrix of $|\hat{x}\rangle$ on $\mathbb{C}^d$.

As a random variable, ρ defines a $\mathbb{U}(d)$-invariant measure on the set of all density operators of rank at most p. By Lemma 12, ρ may be replaced with a random vector variable η and a Haar random U. Due to the permutation invariance and the normalization, we have $\mathbb{E}\eta_i = \mathbb{E}\eta_j = 1/p$, so $\mathbb{E}_\eta \sum_i \eta_i |i\rangle \langle i| = P/p$.

By the convexity of exp,

$$\begin{aligned}
&\mathbb{E}_{x_i} \exp\left[\xi \sum_{i=1}^{2pq} x_i^2\right] \\
&= \mathbb{E}_r \mathbb{E}_\eta \mathbb{E}_U \exp\left[\xi r^2 \operatorname{tr} QU\eta U^\dagger\right] \\
&\geq \mathbb{E}_U \exp\left[\xi (\mathbb{E}_r r^2) \mathbb{E}_\eta \operatorname{tr} QU\eta U^\dagger\right] \\
&= \mathbb{E}_U \exp\left[\xi (dp) \operatorname{tr} QU(P/p)U^\dagger\right].
\end{aligned}$$

we complete the proof. □

Proof of Lemma 6. Recall Markov's inequality: For non-negative real random variable X and $a > 0$, $\Pr[X \geq a] \leq \mathbb{E}X/a$. This is easily seen once we define $Y = a$ if $X \geq a$ and $Y = 0$ if $X < a$, so $Y \leq X$. Then, $\Pr[X \geq a] = \Pr[Y = a] = \mathbb{E}Y/a \leq \mathbb{E}X/a$.

Let us abbreviate $\frac{d}{pq} \operatorname{tr} QUPU^\dagger$ as Z. For any $\xi > 0$ and $z > 0$,

$$\begin{aligned}
\Pr[Z \geq 1 + z] &= \Pr[e^{\xi Z} \geq e^{\xi(1+z)}] \\
&\leq \mathbb{E}_U e^{\xi Z} e^{-\xi(1+z)} \\
&\leq \mathbb{E}_{x_i} \exp\left[\frac{\xi}{pq} \sum_{i=1}^{2pq} x_i^2\right] e^{-\xi(1+z)} \\
&= e^{-\xi(1+z)} \left(1 - \frac{\xi}{pq}\right)^{-pq}
\end{aligned}$$

The last equality is directly evaluated with PDF $\frac{1}{\sqrt{\pi}} e^{-z^2}$. The best bound is when $\xi = pqz/(1+z) > 0$. Substituting this value for ξ, we prove the first inequality in the theorem.

The opposite direction goes similarly. Let $\xi > 0$ and $z \in (0, 1)$.

$$\begin{aligned}
\Pr[Z \leq 1 - z] &= \Pr[e^{-\xi Z} \geq e^{-\xi(1-z)}] \\
&\leq \mathbb{E} e^{-\xi Z} e^{\xi(1-z)} \\
&\leq \mathbb{E} \exp\left[-\frac{\xi}{pq} \sum_{i=1}^{2pq} x_i^2\right] e^{\xi(1-z)} \\
&= e^{\xi(1-z)} \left(1 + \frac{\xi}{pq}\right)^{-pq}
\end{aligned}$$

The best bound is when $\xi = pqz/(1-z) > 0$. Substituting this value for ξ, we prove the second inequality in the theorem.

The last inequality can be proved by examining extreme values of, for example, $g(z) = z - \ln(1+z) - (1 - \ln 2)z^2$. The minimum values in the range $z \in (-1, 1]$ occur at $z = 0, 1$, where $g(z) = 0$. □

[1] Richard Kueng, Holger Rauhut, and Ulrich Terstiege, "Low rank matrix recovery from rank one measurements," (2014), 1410.6913.

[2] Steven T. Flammia and Yi-Kai Liu, "Direct fidelity estimation from few pauli measurements," Phys. Rev. Lett. **106,**, 230501 (2011), 1104.4695.

[3] Christopher A. Fuchs and Jeroen van de Graaf, "Cryptographic distinguishability measures for quantum mechanical states," IEEE Trans. Inf. Theory **45**, 1216 (1999), quant-ph/9712042.

[4] W. K. Wootters, "Statistical distance and hilbert space," Phys. Rev. D **23**, 357–362 (1981).

[5] Michael Nussbaum and Arleta Szkoła, "The chernoff lower bound for symmetric quantum hypothesis testing," The Annals of Statistics **37**, 1040–1057 (2009), quant-ph/0607216.

[6] K. M. R. Audenaert, J. Calsamiglia, R. Muñoz Tapia, E. Bagan, Ll. Masanes, A. Acin, and F. Verstraete, "Discriminating states: The quantum chernoff bound," Phys. Rev. Lett. **98**, 160501 (2007), quant-ph/0610027.

[7] Carl W Helstrom, "Quantum detection and estimation theory," Journal of Statistical Physics **1**, 231–252 (1969).

[8] A.S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*, Publications of the Scuola Normale Superiore (Scuola Normale Superiore, 2011).

[9] E. Bagan, M. Baig, R. Muñoz Tapia, and A. Rodriguez, "Collective versus local measurements in a qubit mixed-state estimation," Phys. Rev. A **69**, 010304 (2004), quant-ph/0307199.

[10] E. Bagan, M. A. Ballester, R. D. Gill, A. Monras, and R. Muñoz Tapia, "Optimal full estimation of qubit mixed states," Phys. Rev. A **73**, 032301 (2006), quant-ph/0510158.

[11] Mădălin Guţă and Jonas Kahn, "Local asymptotic normality for qubit states," Phys. Rev. A **73**, 052108 (2006), quant-ph/0512075.

[12] Mădălin Guţă and Jonas Kahn, "Optimal estimation of qubit states with continuous time measurements," Communications in Mathematical Physics **277**, 127–160 (2008), quant-ph/0608074.

[13] Masahito Hayashi and Keiji Matsumoto, "Asymptotic performance of optimal state estimation in qubit system," Journal of Mathematical Physics **49**, 102101 (2008), quant-ph/0411073.

[14] Jonas Kahn and Mădălin Guţă, "Local asymptotic normality for finite dimensional quantum systems," Communications in Mathematical Physics **289**, 597–652 (2009), 0804.3876.

[15] Masahito Hayashi, *Quantum information: an introduction* (Springer-Verlag, 2006).

[16] M. Keyl, "Quantum state estimation and large deviations," Reveiws in Mathematical Physics **18**, 19–60 (2006), quant-ph/0412053.

[17] Masahito Hayashi, "Asymptotic estimation theory for a finite-dimensional pure state model," Journal of Physics A: Mathematical and General **31**, 4633 (1998), quant-ph/9704041.

[18] Giulio Chiribella, "On quantum estimation, quantum cloning and finite quantum de Finetti theorems," in *Proceedings of the 5th conference on Theory of quantum computation, communication, and cryptography*, TQC'10 (Springer-Verlag, Berlin, Heidelberg, 2011) pp. 9–25, 1010.1875.

[19] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker, and Jens Eisert, "Quantum state tomography via compressed sensing," Phys. Rev. Lett. **105** (2010), 0909.3304.

[20] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity, and efficient estimators," New J. Phys. **14**, 095022 (2012), 1205.2300.

[21] Vladislav Voroninski, "Quantum tomography from few full-rank observables," (2013), 1309.7669.

[22] D. H. Mahler, Lee A. Rozema, Ardavan Darabi, Christopher Ferrie, Robin Blume-Kohout, and A. M. Steinberg, "Adaptive quantum state tomography improves accuracy quadratically," Phys. Rev. Lett. **111**, 183601 (2013), 1303.0436.

[23] Christopher Ferrie and Robin Blume-Kohout, "Minimax quantum tomography: the ultimate bounds on accuracy," (2015), 1503.03100.

[24] Richard D. Gill and Serge Massar, "State estimation for large ensembles," Phys. Rev. A **61**, 042312 (2000), quant-ph/9902063.

[25] Masahito Hayashi, "Quantum estimation and the quantum central limit theorem," American Mathematical Society Translations 2, **277**, 99–123 (2009), quant-ph/0608198.

[26] M. Keyl and R. F. Werner, "Estimating the spectrum of a density operator," Phys. Rev. A **64**, 052311 (2001), quant-ph/0102027.

[27] Masahito Hayashi and Keiji Matsumoto, "Quantum universal variable-length source coding," Phys. Rev. A **66**, 022311 (2002), quant-ph/0202001.

[28] Matthias Christandl and Graeme Mitchison, "The spectra of quantum states and the kronecker coefficients of the symmetric group," Commun. Math. Phys. **261**, 789–797 (2006).

[29] A.M. Childs, A. W. Harrow, and P. Wocjan, "Weak Fourier-Schur sampling, the hidden subgroup problem, and the quantum collision problem," in *Proc. of STACS*, LNCS, Vol. 4393 (2007) pp. 598–609, quant-ph/0609110.

[30] Ryan O'Donnell and John Wright, "Quantum spectrum testing," in *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, STOC '15 (2015) pp. 529–538, 1501.05028.

[31] Richard D. Gill, "Conciliation of bayes and pointwise quantum state estimation: Asymptotic information bounds in quantum statistics," (2005), math/0512443.

[32] Fuyuhiko Tanaka, "Quantum minimax theorem," (2014), 1410.3639.

[33] VP Belavkin, "Optimum distinction of non-orthogonal quantum signals," Radio Engineering and Electronic Physics **20**, 39–47 (1975).

[34] VP Belavkin, "Optimal multiple quantum statistical hypothesis testing," Stochastics: An International Journal of Probability and Stochastic Processes **1**, 315–345 (1975).

[35] Paul Hausladen and William K. Wootters, "A 'pretty good' measurement for distinguishing quantum states," Journal of Modern Optics **41**, 2385–2390 (1994).

[36] H. Barnum and E. Knill, "Reversing quantum dynamics with near-optimal quantum and classical fidelity," J. Math. Phys. **43**, 2097–2106 (2002), quant-ph/0004088.

[37] A. W. Harrow and A. J. Winter, "How many copies are needed for state discrimination?" IEEE Trans. Inf. Theory **58**, 1–2 (2012), quant-ph/0606131.

[38] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," IEEE Trans. Inf. Theory **48**, 569–579 (2002), quant-ph/0012127.

[39] Ryan O'Donnell and John Wright, "Efficient quantum tomography," (2015), 1508.01907.

[40] Jonas Kahn and M. Guţă, "Quantum stochastics and information: Statistics, filtering and control," (World Scientific, 2008) Chap. Local asymptotic normality and optimal estimation for d-dimensional quantum systems, pp. 300–322.

[41] William Fulton and Joe Harris, *Representation Theory: A first course*, Graduate Texts in Mathematics, Vol. 129 (Springer, 2004).

[42] Necdet Batir, "Inequalities for the gamma function," Archiv der Mathematik **91**, 554–563 (2008).

[43] Jon Tyson, "Error rates of belavkin weighted quantum measurements and a converse to holevo's asymptotic optimality theorem," Phys. Rev. A **79**, 032343 (2009), 0907.1884.

[44] A. S. Holevo, "Bounds for the quantity of information transmitted by a quantum communication channel," Problems of Information Transmission **9**, 177–183 (1973).

[45] Robert M Fano, *The transmission of information* (M.I.T. Press and John Wiley and Sons, New York and London, 1961).

[46] Stanislaw J. Szarek, "Nets of grassmann manifold and orthogonal group," in *Proceedings of Research Workshop on Banach Space Theory*, edited by Bor-Luh Lin (The University of Iowa, 1981) pp. 169–185.

[47] StanislawJ. Szarek, "The finite dimensional basis problem with an appendix on nets of grassmann manifolds," Acta Mathematica **151**, 153–179 (1983).

[48] A Winter, "Quantum and classical message identification via quantum channels," Quantum Inf. Comput. **4**, 563–578 (2004), quant-ph/0401060.

[49] Troy Lee, Ignacio Villanueva, Zhaohui Wei, and Ronald de Wolf, In Preparation (2015).

[50] Patrick Hayden, Debbie W. Leung, and Andreas Winter, "Aspects of generic entanglement," Commun. Math. Phys. **265**, 95–117 (2006), quant-ph/0407049.

[51] A. Winter, "Compression of sources of probability distributions and density operators," (2002), quant-ph/0208131.

[52] D. Bacon, I. L. Chuang, and A. W. Harrow, "The quantum Schur and Clebsch-Gordan transforms: I. Efficient qudit circuits," in *Proc. of SODA* (2007) pp. 1235–1244, quant-ph/0601001.

[53] R. Beals, "Quantum computation of Fourier transforms over symmetric groups," in *Proceedings of the 29th Annual ACM Symposium on the Theory of Computation (STOC)* (ACM Press, El Paso, Texas, 1997) pp. 48–53.

[54] Aram W. Harrow, *Applications of coherent classical communication and Schur duality to quantum information theory*, Ph.D. thesis, M.I.T., Cambridge, MA (2005), quant-ph/0512255.

[55] Raban Iten, Roger Colbeck, Ivan Kukuljan, Jonathan Home, and Matthias Christandl, "Quantum circuits for isometries," (2015), 1501.06911.

[56] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost, "Quantum principal component analysis," Nature Physics **10**, 631–633 (2014).

[57] Patrick Hayden, Debbie Leung, Peter W. Shor, and Andreas Winter, "Randomizing quantum states: Constructions and applications," Commun. Math. Phys. **250**, 371–391 (2004), quant-ph/0307104.