Learning Quantum States Without Entangled Measurements

by

Angus Lowe

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Combinatorics and Optimization (Quantum Information)

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

How many samples of a quantum state are required to learn a complete description of it? As we will see in this thesis, the fine-grained answer depends on the measurements available to the learner, but in general it is at least $\Omega(d^2/\epsilon^2)$ where d is the dimension of the state and ϵ the trace distance accuracy. Optimal algorithms for this task – known as quantum state tomography – make use of powerful, yet highly impractical *entangled measurements*, where some joint measurement is performed on all copies of the state. What can be accomplished without such measurements, where one must perform measurements on individual copies of the states?

In Chapter 2 we show a relationship between the recently proposed quantum online learning framework and quantum state tomography. Specifically, we show that tomography can be accomplished using online learning algorithms in a black-box manner and $\tilde{O}(d^4/\epsilon^4)$ two-outcome measurements on separate copies of the state. The interpretation of this approach is that the experimentalist uses informative measurements to teach the learner by helping it make "mistakes" on measurements as early as possible.

We move on to proving lower bounds on tomography in Chapter 3. First, we review a known lower bound for entangled measurements as well as a $\Omega(d^3/\epsilon^2)$ lower bound in the setting where non-entangled measurements are made non-adaptively, both due to Ref. [18]. We then derive a novel bound of $\Omega(d^4/\epsilon^2)$ samples when the learner is further restricted to observing a constant number of outcomes (e.g., two-outcome measurements). This implies that the folklore "Pauli tomography" algorithm is optimal in this setting.

Understanding the power of *adaptive measurements*, where measurement choices can depend on previous outcomes, is currently an open problem. In Chapter 4 we present two scenarios in which adapting on previous outcomes makes *no difference* to the number of samples required. In the first, the learner is limited to adapting on at most $o(d^2/\epsilon^2)$ of the previous outcomes. In the second, measurements are drawn from some set of at most $\exp(O(d))$ measurements. In particular, this second lower bound implies that adaptivity makes no difference in the regime of efficiently implementable measurements, in the context of quantum computing.

Finally, we apply the above technique to the problems of classical shadows and shadow tomography to obtain similar lower bounds. Here, one is interested only in determining the expectations of some fixed set of observables. We once again find that, for the worst-case input of observables, adaptivity makes no difference to the sample complexity when considering efficient, non-entangled measurements. As a corollary, we find a straightforward algorithm for shadow tomography is optimal in this setting.

Acknowledgements

I am fortunate to have been advised by Ashwin Nayak, whose patience and dependability enabled me to complete this thesis in uncertain times. Much of the progress I have made toward becoming a better researcher I owe to our technical discussions, as well as his general advice to avoid pattern recognition in favour of deep understanding. I hope to have conveyed something befitting of this suggestion somewhere in my thesis. I would also like to express my gratitude to Vern Paulsen and David Gosset, for generously agreeing to read my thesis and bearing with me as I made sure things were in order.

I would like to thank Sitan Chen for a helpful correspondence via email, where he pointed out the tools necessary to prove Lemma 4.1.4, as well as Hsin-Yuan Huang for pointing out lower bounds for unentangled classical shadows in Ref. [22].

I am thankful to have had the support of many people from the University of Waterloo. Thank you to my classmates and lecturers for classes and discussions on many interesting topics, for teaching me things I didn't know I didn't know. Thanks in particular to Shlok Nahar, John Bostanci, and Alex Kerzner for helpful discussions related to this thesis. I am also grateful to Melissa Cambridge, for guiding me through the process of completing a Master's degree.

To my partner, Uttara, thank you for being my best friend and the source of so much of my happiness.

Finally, I am forever grateful to my family for their unconditional love and support. Without them I would not have the opportunity to study what I am passionate about. And thank you to Aegon, for being such a good boy.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

This thesis is concerned with how many samples of a quantum state ρ are required to learn some description of it, given various restrictions on what can be done with these samples. A *sample* in this context amounts to preparing the state ρ in some register. For the most part, we focus on quantum state tomography, the fundamental task of estimating ρ to within some known accuracy ϵ in the standard trace distance between states. We will be especially interested in how the number of samples scales with the dimension d of the state to be learned, since this quantity grows exponentially with the number of qubits comprising the system and is therefore, in theory, the primary obstacle toward accomplishing the task.

In the setup where samples of ρ are prepared on registers that may be simultaneously measured, it is said that the measurements are *entangled*. A series of breakthrough works [31, 18, 39] proved that $O(d^2/\epsilon^2)$ samples suffice to perform tomography using entangled measurements, matching an information-theoretic lower bound due to Ref. [18] and improving upon previous upper bounds by a factor of d. Note that throughout this thesis, we consider the case where the states can be full-rank in the interest of worst-case bounds. We say that the *sample complexity* of entangled quantum state tomography is $O(d^2/\epsilon^2)$ (and $\Omega(d^2/\epsilon^2)$, by the lower bound).

From a practical standpoint, however, entangled measurements are problematic for a few reasons. Firstly, in the case where one has access to just a single register that can be prepared in the state ρ, entangled measurements are impossible. (For instance, one might wish to perform tomography on the output state of a quantum computer by repeating a computation). Second, even if one had access to n copies of the state simultaneously, the entangled measurements may require that they undergo some joint unitary evolution

(hence the term "entangled") which is not possible if, for instance, one has n non-entangled quantum systems. Finally, it is not clear how to implement the arbitrary nd-dimensional measurements which arise in the entangled case, even with access to a suitably large system that can be prepared in the state $\rho^{\otimes n}$. For these reasons, there is strong motivation to consider restricted measurement models which have been coined *unentangled* measurements by some sources [39, 13].

Within the unentangled 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 state is discarded. This means that the number of samples is equal to the number of measurements performed. Results pertaining to the sample complexity of unentangled tomography predate those for entangled measurements, potentially due to the practical relevancy of this case. Two prominent examples are the folklore "Pauli tomography algorithm" (outlined in Section 8.4.2 in Nielsen and Chuang [30]) and an algorithm due to Kueng, Rauhut, and Terstiege [26] (KRT) based on low-rank matrix recovery. In both examples, the upper bound on the sample complexity is worse than in the entangled case. However, it is unclear whether this reflects a fundamental limit on what can be achieved without entangled measurements.

In Chapter 2, we give a meta-algorithm for quantum tomography using unentangled, nonadaptive binary measurements using at most $\tilde{O}(d^4/\epsilon^4)$ samples. In terms of the dimension, this matches an information-theoretic lower bound we prove for this setting in Chapter 3. We use the term meta-algorithm to denote that it is based on using the recently proposed quantum online learning framework [4] in a black-box manner. In other words, the procedure is agnostic to the way in which one chooses to implement the online learning algorithm, so long as it has a sufficiently small "mistake bound". The techniques we use to analyze this algorithm's sample complexity involve an anti-concentration result to argue that the mistake bound is likely to be saturated using not too many measurements. For this, it suffices to consider rotating binary measurement operators according to unitary 4-designs. This is in direct analogy with the Hamiltonian updates procedure due to Brandão, Kueng, and França [12]. Here, an upper bound of $\tilde{O}(d^3/\epsilon^4)$ is shown using random d-outcome measurements and an adaptation of the matrix-exponentiated gradient method for online learning which applies in the d-outcome case [35].

What is the best possible sample complexity using unentangled measurements? Haah et al. [18] partially resolve this question by providing a $\Omega(d^3/\epsilon^2)$ lower bound which matches the upper bound in the KRT protocol, under the assumption that the choices of measurement are independent of any previous outcomes (referred to as *nonadaptive measurements*). However, their bound is not tight for the simplest of tomography protcols (such as the Pauli tomography algorithm), and does not exhaust a particularly extensive set of realizable mea-

	Nonadaptive		$o(d^2/\epsilon^2)$-adaptive	Adaptive+efficient	
# outcomes	$O(1)$	arbitrary	arbitrary	$O(1)$	arbitrary
Upper bound	$O(d^4/\epsilon^2)$	$O(d^3/\epsilon^2)$ [26]	$O(d^3/\epsilon^2)$	$O(d^4/\epsilon^2)$	$O(d^3/\epsilon^2)$
Lower bound	$\Omega(d^4/\epsilon^2)$ [$*$]	$\Omega(d^3/\epsilon^2)$ [18]	$\Omega(d^3/\epsilon^2)$ [$*$]	$\tilde{\Omega}(d^4/\epsilon^2)$ [$*$]	$\Omega(d^3/\epsilon^2)$ [$*$]

Table 1.1: Best known upper and lower bounds for the sample complexity of unentangled quantum state tomography. $\tilde{\Omega}$ hides $\log(d)$ factors, lack of citation indicates folklore or implied by other bounds, and [$*$] denotes results from this thesis. For more on the assumptions corresponding to the final two columns, see Sections 4.1 and 4.2, respectively.

surement strategies. Indeed, numerous tomography algorithms have been proposed [24, 27, 41] which utilize unentangled *adaptive measurements* where measurements can depend on previous outcomes. Chapters 3 and 4 of this thesis are dedicated to providing lower bounds in each of these measurement scenarios, and in particular in Chapter 4 we derive lower bounds robust to a wide class of adaptive measurements. We summarize our lower bounds in comparison to previous work in Table 1.1.

In Chapter 3 we begin by describing the basic framework for proving lower bounds on the task of quantum tomography. We make use of the observation that state discrimination of sufficiently well-separated states reduces to tomography with sufficient accuracy. Our lower bounds then follow from difficult instances of the state discrimination problem, where 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. [36].) To the best of the author's knowledge, the method was first employed in the context of tomography by Flammia et al. [16] to derive a $\Omega(d^4/\log(d))$ lower bound when one is restricted to using adaptive binary Pauli measurements. The general method has since been used successfully in the lower bounds for nonadaptive measurements due to Haah et al. [18], as well as for lower bounds of a different learning problem known as *classical shadows* [23], where the measurements are once again assumed to be nonadaptive.

We first review the proof of the lower bound due to Haah et al. [18], and in the process improve it by a factor of d when the measurements are restricted to having a constant number of outcomes. This implies that the straightforward Pauli tomography algorithm – which uses two-outcome measurements – is optimal in this case. Our analysis leverages a known connection between the mutual information of two random variables and the χ^2 divergence of their distributions, as well as techniques for Haar integration based on symmetry. Additionally, our proof does not require that the measurements be rank-one

POVMs as in Ref. [18].

In Chapter 4 we shift our attention to the case where the sequence of measurements can be made adaptively, which represents an intermediate restriction between nonadaptive and entangled measurements. Much less is known about the sample complexity of learning quantum states using adaptive measurements, though it has arguably more significance: proving advantages for entangled measurements over unentangled measurements for some learning task amounts to showing that adaptive measurements have strictly worse sample complexity. It was posed as an open problem in Wright [39] to provide examples where this is the case, and there has been some recent progress here, though not for the problem of tomography. As recently as 2020, Bubeck, Chen, and Li [13] gave the first unconditional separation between entangled vs. unentangled measurements, for the problem of quantum state certification. Following this, Huang, Kueng and Preskill [22] proved an exponential separation for the problem of determining the expectations of Pauli operators to constant accuracy.

The lower bounds we present in Chapter 4 are the first to show a setting in which adaptivity makes *no difference* to the worst-case sample complexity of learning a quantum state. The catch is that we assume additional restrictions on the measurements beyond belonging to the class of unentangled measurement schemes. However, we believe these assumptions to be fairly mild, and the author is not aware of a proposal for adaptive quantum state tomography which fails to meet at least one of these assumptions.

The first restriction admitting a tight lower bound is one we term *limited adaptivity*, where the learner has infinitely many measurement settings, but may only adapt on a fixed subset of the outcomes which is not larger than $\Omega(d^2/\epsilon^2)$. The techniques we use to obtain this bound are the same as the ones employed in the unconditional lower bounds of Ref. [13].

The second kind of restriction under which we can show tight lower bounds is that of efficiently implementable measurements. This comes from the fact that the lower bounds we prove are robust to any measurement scheme which uses adaptive measurements drawn from a fixed set of up to $\exp(O(d))$ settings. We explain this in greater detail in Chapter 4. We arrive at this lower bound by adversarially constructing our instance of the state discrimination problem to be as difficult as possible for the specific set of measurements under consideration.

Finally, we apply this technique to obtain a lower bound for the problem of classical shadows robust to adaptively chosen measurements, so long as they are efficiently implementable. Here one is interested in estimating the expectations of some collection of observables, with practical applications ranging from entanglement verification to near-term proposals of variational quantum algorithms [23, 33]. We also find that a simpler procedure

than the one given in Ref. [23] is sample-optimal for classical shadows with unentangled, efficient measurements, and under worst-case assumptions for the set of observables.

1.2 Preliminaries

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

Sets

Let $\mathbb{Z}_+$ denote the set of nonnegative integers, $\mathbb{U}(d)$ 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 $\mathsf{H}(d)$ which is positive semidefinite, and $\mathsf{D}(d)$ the subset of $\mathsf{Psd}(d)$ which has unit trace (i.e., the set of d-dimensional quantum states). For some positive integer $N > 0$, let $[N] = \{1, \ldots, N\}$.

Operators

For any square operator $A \in \mathbb{C}^{d\times d}$ let $A^\dagger$ denote its adjoint. Let $X \in \mathsf{H}(d)$ be a Hermitian operator with spectral decomposition $\sum_{i=1}^d \lambda_k(X)|k\rangle\langle k|$, where $\lambda_d(X) \leq \cdots \leq \lambda_1(X)$ are its eigenvalues. For a function of the form $f : \mathbb{C} \to \mathbb{C}$, we define $f(X) = \sum_{i=1}^d f(\lambda_k(X))|k\rangle\langle k|$. Let $\|A\|_1 = \mathrm{Tr}(\sqrt{A^\dagger A})$ denote the "trace norm" of the operator A and note that $\|X\|_1 = \sum_{k=1}^d |\lambda_k(X)|$. Let $\|A\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^\dagger A)}$ be the Frobenius norm of the operator A and note that $\|X\|_{\mathrm{F}}^2 = \sum_{k=1}^d |\lambda_k(X)|^2$. Let $\|A\|$ be the spectral norm of the operator A which 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 have the relation $P \preceq Q$ if and only if $Q - P \in \mathsf{Psd}(d)$. A useful result is that, for any $\rho, \sigma \in \mathsf{D}(d)$ we have

$$\|\rho - \sigma\|_1 = 2 \max_{0\preceq P\preceq \mathbb{1}} \mathrm{Tr}(P(\rho - \sigma))$$

where the *trace distance* between two quantum states ρ, σ is defined to be the quantity on the left-hand side. Let $A, B \in \mathsf{H}(d)$ and consider the operator $A \otimes B$. We denote by $\mathrm{Tr}_2(\cdot)$ the operation of tracing out the second system, i.e., $\mathrm{Tr}_2(A \otimes B) = A\mathrm{Tr}(B)$.

Permutation operator

The swap operator W acting on $(\mathbb{C}^d)^{\otimes 2}$ is 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 operator W_π for each $\pi \in S_n$ and acting on $(\mathbb{C}^d)^{\otimes n}$ by the action

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

for every choice of vectors $|x_1\rangle, \ldots, |x_n\rangle \in \mathbb{C}^d$.

Random variables

We denote random variables using bold font, including matrix-valued random variables. We use lowercase 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 $\mathbb{E}_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{y}} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y})$.

1.3 Facts from information theory

Classical information theory

First, let us consider discrete random variables taking values on the same space. We 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

space is

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

(Throughout this thesis, log denotes logarithm base 2.) We next define some entropic quantities. Let $\boldsymbol{x}$ be a discrete random variable taking values in $\mathcal{X}$ with distribution $p_{\boldsymbol{x}}$. 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 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 1.3.1 (Mutual information)**.** Consider two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ where $\boldsymbol{x}$ has marginal distribution $p_{\boldsymbol{x}}$ and $\boldsymbol{y}$ is discrete. Let $p_{\boldsymbol{y}|x}$ be the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x} = x$ and $p_{\boldsymbol{y}}$ the marginal distribution of $\boldsymbol{y}$. The *mutual information* between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

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

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,

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

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

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

where $(\boldsymbol{x}|\boldsymbol{z} = z)$ is the marginal distribution of $\boldsymbol{x}$, 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

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

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 (to be covered later in this section), in the case where there is some restriction on the measurements.

Fact 1.3.2. *Let $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ be random variables, and suppose that $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent given $\boldsymbol{x}$. Then*

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

Fact 1.3.3 (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 1.3.4 (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 the conditional distribution of $\boldsymbol{z}$ depends only on $\boldsymbol{y}$ and is conditionally independent of $\boldsymbol{x}$ (Ref. [14], Section 2.8). Under this assumption, the following lemma holds, which is indispensable toward proving information-theoretic lower bounds.

Lemma 1.3.5 (Fano's inequality [15])**.** *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 $p_e := \Pr[\boldsymbol{x} \neq \hat{\boldsymbol{x}}]$, and $H(\cdot)$ is the binary entropy function.

Corollary 1.3.6. *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 $\boldsymbol{x}$ which is uniformly random over N distinct values, 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 1.3.5. By Lemma 1.3.5 we have $I(\boldsymbol{x} : \boldsymbol{y}) \geq H(\boldsymbol{x}) - p_e \log(N) - H(p_e)$. Using the fact that $H(\boldsymbol{x}) = \log(N)$ for uniformly random $\boldsymbol{x}$ and $H(p_e) \leq 1$ we obtain $I(\boldsymbol{x} : \boldsymbol{y}) \geq (1 - p_e) \log(N) - 1$. □

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

Definition 1.3.7 (χ^2-divergence)**.** The χ^2-divergence between two discrete distributions $p, q : \mathcal{X} \to [0, 1]$ defined on the same space $\mathcal{X}$ is

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

These divergences are related in the following way.

Lemma 1.3.8 (KL vs. χ^2 inequality)**.** *Let $p, q : \mathcal{X} \to [0,1]$ be discrete distributions defined on the same space $\mathcal{X}$. We have*

$$D_{\mathrm{KL}}(p \parallel q) \leq \log(e)\ \chi^2(p \parallel q).$$

See Ref. [32] for example.

Quantum information theory

We now move on to some elementary definitions and results from quantum information theory.

Definition 1.3.9 (Ensemble)**.** Consider N mixed states $\rho_1, \ldots, \rho_N$ along with some random variable $\boldsymbol{x}$ taking values in $[N]$. We refer to the random state $\rho_{\boldsymbol{x}}$ as an *ensemble* of states.

The *von Neumann entropy* of the quantum state $\rho \in \mathsf{D}(d)$ is defined as

$$S(\rho) := -\mathrm{Tr}(\rho \log(\rho)) = -\sum_{i=1}^{d} \lambda_i(\rho) \log(\lambda_i(\rho)).$$

Theorem 1.3.10 (Holevo's theorem, Theorem 12.1 in [30])**.** *Consider N mixed states $\rho_1, \ldots, \rho_N$ and let $\boldsymbol{x}$ be some random variable distributed over $[N]$ according to $p_{\boldsymbol{x}}$. Let $\boldsymbol{y}$ be the random variable corresponding to the outcome obtained from performing some measurement on $\rho_{\boldsymbol{x}}$, and define the state $\rho := \sum_{x=1}^{N} p_{\boldsymbol{x}}(x)\rho_x$. It holds that*

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq S(\rho) - \sum_{x=1}^{N} p_{\boldsymbol{x}}(x) S(\rho_x).$$

The quantity in the right-hand side of the above inequality is sometimes referred to as the *Holevo information* of the ensemble $\rho_{\boldsymbol{x}}$.

1.4 Measurement models for learning quantum states

In general, a quantum measurement of a d-dimensional quantum state is described by a positive operator-valued measure (POVM) $\mathcal{M}$ mapping quantum states to diagonal operators,

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

where $\{M_z\}_z \subset \mathsf{Psd}(d)$ is a set of *measurement operators* satisfying $\sum_z M_z = \mathbb{1}$. We focus on measurements with a finite number of outcomes throughout this thesis. The distribution of the random outcome $\boldsymbol{z}$ from measuring the state ρ is described by the PMF $p_z = \mathrm{diag}(\mathcal{M}(\rho))$, so that $p_{\boldsymbol{z}}(z) = \mathrm{Tr}(M_z\rho)$ for all outcomes z. We describe next the three models of measurement which we will be adopting in this thesis. Throughout, we assume that the task is to learn properties of some unknown d-dimensional state $\rho \in \mathsf{D}(d)$ using as few samples of ρ as possible. As mentioned in Section 1.1, a *sample* refers to the act of preparing a register in the state ρ for measurement by the learner.

1.4.1 Entangled measurements

In the most general model, the learner is provided a register in the d^n-dimensional state $\rho^{\otimes n}$ so that the number of samples is n, and an arbitrary measurement of this product state is performed. The task is then to use the outcome of this measurement to infer properties of ρ.

State discrimination

A classic example is that of quantum state discrimination, where the goal is to identify the state ρ_i picked from a known set of alternatives $\{\rho_j\}$ with the highest possible success probability. The discrimination is enabled by the fact that taking the tensor product effectively amplifies the trace distance between the alternatives, so that for sufficiently large n it becomes possible to distinguish them via some measurement. (See Ref. [29] for an upper bound on n.) The intuition that for large enough n there exists a measurement which identifies the state with high probability continues to hold for the task of quantum tomography with finite precision, in which case the set of alternatives is $\mathsf{D}(d)$, but the success criteria is relaxed to outputting states that are close enough to the true state. In fact, sample-optimal tomography in this setting is possible using, roughly speaking, a measurement analogous to the optimal measurement for state discrimination [31, 18]. The connection with state discrimination is also made apparent by observing that discrimination reduces to sufficiently accurate tomography of the states, and we will return to this point in Chapter 3 where we derive lower bounds for tomography in each of our measurement models.

1.4.2 Unentangled measurements

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. The class of measurements corresponding to this scenario is known as *unentangled*, where the number of samples used is equivalent to the number of measurements performed. We describe the two main models of measurement which fall into this class below.

Nonadaptive measurements

Consider n copies of the state ρ prepared in the above manner, so that they must be measured individually. In the *nonadaptive* measurement model, we use a sequence of d-dimensional measurements $\mathcal{M}_i$ for $i = 1, \ldots, n$ which are determined beforehand. Equivalently, we measure the state $\rho^{\otimes n}$ using the *product measurement* $\mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$. 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 then be incorporated into the measurement itself i.e., the resulting mapping is still some fixed measurement.

Adaptive measurements

In the *adaptive* measurement model, the choice of each d-dimensional measurement in the sequence can depend on the previous outcomes obtained. 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 measurements. For each possible value of $y_{<i}$ there is a set of measurement operators $\{M_{y_i}^{y_{<i}}\}_{y_i}$ corresponding to the different outcomes y_i of the i^{th} measurement, such that the measurement has the action

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

for any state $\rho \in \mathsf{D}(d)$.

1.5 Upper bounds on quantum tomography

1.5.1 $O(d^2/\epsilon^2)$-tomography from entangled measurements

In the entangled measurement model, it has been shown by O'Donnell and Wright [31] and Haah et al. [18] that $O(d^2/\epsilon^2)$ copies of the state suffice to estimate it to ϵ-accuracy

in trace distance with high probability[1]. At the same time, a matching lower bound was also shown in [18] meaning that the sample complexity of tomography in the entangled measurement setting is essentially solved. A full description of these algorithms is outside the scope of this thesis, 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 thesis of Wright [39].

1.5.2 $O(d^3/\epsilon^2)$-tomography from random basis measurements

In the remainder of this thesis we will make use of unitary t-designs, which we define below.

Definition 1.5.1 (Unitary t-design)**.** We refer to the random unitary operator $\boldsymbol{V}$ taking values in $\mathbb{U}(d)$ as a *unitary t-design* (or say that it *comprises* a unitary t-design) if the following holds for every operator $X \in (\mathbb{C}^{d\times d})^{\otimes t}$:

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

where μ is the Haar measure on the space of d-dimensional unitary operators. (See Appendix A for more on Haar integration.)

We now describe an algorithm which achieves a sample complexity of $O(d^3/\epsilon^2)$ for ϵ-accurate tomography (in trace distance) using independent, rank-one measurements. Following a matching lower bound due to Haah et al. [18], this algorithm is sample-optimal in the nonadaptive measurement setting. The analysis we present is due to Wright [39] (Section 5.1), with minor differences and pointing out that unitary 2-designs are sufficient. A more detailed analysis in the low-rank case is originally due to [26].

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 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 i.i.d. 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 and $\boldsymbol{j}_i$ is the outcome from the i^{th} measurement. Let $\hat{\rho}(U, j) := (d+1)U|j\rangle\langle j|U^\dagger - \mathbb{1}$ for any $U \in \mathbb{U}(d)$ and $j \in [d]$.

Proposition 1.5.2. *It holds that*

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

[1]Originally, the upper bound presented in Haah et al. [18] had an additional factor of $\log(d/\epsilon)$, which was subsequently removed in the thesis of Wright [39].

Proof. We defer the calculation of some Haar integrals to Appendix A. Let $p_{\boldsymbol{U}}$ denote the distribution of $\boldsymbol{U}$ and $p_{\boldsymbol{j}|U}(j)$ the probability of obtaining outcome j given that U is drawn. We have

$$\begin{aligned}\mathbb{E}\,\boldsymbol{U}|\boldsymbol{j}\rangle\langle\boldsymbol{j}|\boldsymbol{U}^\dagger &= \sum_{j=1}^{d} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} p_{\boldsymbol{j}|\boldsymbol{U}}(j)\ \boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger \\ &= \sum_{j=1}^{d} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} \langle j|\boldsymbol{U}\rho\boldsymbol{U}^\dagger|j\rangle\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger. \end{aligned} \tag{1.1}$$

Consider a specific term j in the sum above. We may write that term equivalently as

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim p_{\boldsymbol{U}}} \mathrm{Tr}_2\left((\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger)^{\otimes 2}(\mathbb{1}\otimes\rho)\right) = \mathrm{Tr}_2\left(\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}}(\boldsymbol{U}|j\rangle\langle j|\boldsymbol{U}^\dagger)^{\otimes 2}(\mathbb{1}\otimes\rho)\right) \tag{1.2}$$

where the equality follows from linearity of trace and the fact that $\boldsymbol{U}$ is a 2-design. The first relation in Proposition A.0.2 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

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

Substituting into the right-hand side of Eq. (1.2) 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 fact that this holds for any $j \in [d]$ and substituting into (1.1) we obtain the relation $\mathbb{E}\,\boldsymbol{U}|\boldsymbol{j}\rangle\langle\boldsymbol{j}|\boldsymbol{U}^\dagger = \frac{1}{d+1}(\mathbb{1}+\rho)$. 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. 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

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

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

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

where the inequality used $\operatorname{Tr}(\rho^2) \geq 0$ and the final line comes from the following fact: for a Hermitian matrix A, we have $\operatorname{Tr}(A^2) = \sum_{i=1}^d \lambda_i(A)^2$ and in this case, the operator $(d+1)U|j\rangle\langle j|U^\dagger - \mathbb{1}$ has eigenvalues which are all -1 except for one eigenvalue, which 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 $n = O(d^3/\epsilon^2)$ gives us the desired upper bound in expectation, and we can easily convert this convergence in expectation into convergence with high probability using Markov's inequality.

1.5.3 $O(d^4/\epsilon^2)$-tomography from 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 will be proving in Chapter 3 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

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

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 $\boldsymbol{z}_i \in \{\pm 1\}$ defined in the obvious way. Then $\boldsymbol{z}_i$ is an unbiased estimator for the i^{th} Pauli coefficient, and performing this measurement $s \in \mathbb{Z}_+$ times results in i.i.d. random variables $\{\boldsymbol{z}_{i,j}\}_{j=1}^s$. Let us then take the empirical average of the s samples corresponding to the i^{th} Pauli measurement $\boldsymbol{\mu}_i := \frac{1}{s}\sum_{j=1}^s \boldsymbol{z}_{i,j}$, 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

$$\begin{aligned}
\mathbb{E}\,\|\hat{\boldsymbol{\rho}} - \rho\|_{\mathrm{F}}^2 &= \frac{1}{d}\sum_{i=1}^{d^2} \mathbb{E}\,|\boldsymbol{\mu}_i - \mathrm{Tr}(P_i\rho)|^2 \\
&= \frac{1}{d}\sum_{i=1}^{d^2} \mathrm{Var}[\boldsymbol{\mu}_i] \\
&= \frac{1}{ds^2}\sum_{i=1}^{d^2}\sum_{j=1}^{s} \mathrm{Var}[\boldsymbol{z}_{i,j}] \\
&\leq \frac{d}{s}
\end{aligned}$$

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.

Chapter 2

Online learning implies efficient tomography

The number of measurements required in any procedure for tomography of an n-qubit quantum state must grow exponentially in n, which we will show in the beginning of Chapter 3. This apparent difficulty, however, can be misleading depending on what the learner is ultimately interested in, and so several alternative models have been proposed with relaxed criteria for what constitutes learning the state. Of particular interest is mistake-bounded quantum online learning [4]. Here, the learner undergoes T rounds of communication with an adversary. In each round, the adversary provides the learner with a binary measurement, for which the learner must output an estimate of the corresponding expectation value. At the end of each round, the adversary reveals as feedback an approximation of the true expectation, and the learner is considered to have made a "mistake" if their estimate differed too much from the feedback. The goal of the learner is to output a sequence of estimates such that, after M mistakes have been made, the learner is correct (not mistaken) on *all* future measurements provided by the adversary. Somewhat surprisingly, in Ref. [4] it was shown that there exist strategies for which M is linear in n.

It is clear that performing full state tomography on the unknown state beforehand would allow one to succeed in this online learning model. But what about the other direction – is there a sense in which the ability to perform mistake-bounded online learning allows one to successfully perform tomography? In this section, we show that the answer to this question is yes with the additional guarantee that such a strategy can be made nearly optimal in terms of the sample complexity, in the setting of nonadaptive measurements. In other words, online learning algorithms such as the recently proposed RSOA [10], and FTPL (see [20] for example), or MMW [35, 8] can be used in a black-box manner to perform

tomography using a number of measurements which scales nearly optimally.

This result improves upon the sample complexity guarantees originally given in Ref. [40] for using an online learning algorithm to perform tomography, where there are no strong claims about the number of copies of the state required. The techniques used in this section are in close analogy with the *Hamiltonian updates* tomography protocol of Franca et al. [12] which uses d-outcome measurements, and the result in this chapter can be interpreted as an extension of that algorithm to the case of binary measurements, which are of particular relevance to the framework of online learning of quantum states. Generalizing the online learning results which are known about the binary case to more measurement outcomes is an open problem proposed in Ref. [4].

2.1 Quantum online learning

Let $\epsilon \in (0, 1)$ be a *mistake parameter* and consider a *true state* $\rho \in \mathsf{D}(d)$ unknown to the learner. In the setting of online learning of quantum states, we consider two parties – the online learner and the adversary – who undergo a sequence of iterations (or rounds) $t = 1, 2, 3, \ldots$ in which the online learner constructs an estimate $\omega_t \in \mathsf{D}(d)$ of the true state ρ. At the end of each round the learner receives from the adversary a measurement operator $0 \preceq E_t \preceq \mathbb{1}$ as well as feedback y_t satisfying $|y_t - \mathrm{Tr}(E_t\rho)| \leq \epsilon/3$. We then say that the learner suffers a loss, as defined below.

Definition 2.1.1 (Loss)**.** The *loss* incurred by the learner in round t is $\ell_t(\mathrm{Tr}(E_t\omega_t))$, where the *loss function* $\ell_t : [0, 1] \to \mathbb{R}$ is defined by

$$\ell_t(x) = |x - y_t|.$$

The loss function indicates how well the current estimate of the state predicts the expectation of E_t. One may consider different expressions for the loss function, but we restrict ourselves to what is referred to as the L_1 loss here. We sometimes write ℓ_t as shorthand for the loss in round t – which is $\ell_t(\mathrm{Tr}(E_t\omega_t))$ – when it should be clear from context this is the quantity of interest.

We say that a "mistake" has been made in iteration t if $|\mathrm{Tr}(E_t\omega_t) - \mathrm{Tr}(E_t\rho)| > \epsilon$, and our goal is to provide an upper bound on the number of iterations where the online learner makes a mistake, which we denote by M. We will show that there is a strategy according to which the online learner makes at most $M = O(\log(d)/\epsilon^2)$ mistakes, regardless of the sequence of measurements presented to them. Furthermore, the strategy which achieves this is the one which minimizes the excess total loss over the best strategy in hindsight, a quantity which is captured in the following definition.

Definition 2.1.2 (Regret)**.** Let $T > 0$ be a positive integer and ω_t be a sequence of estimates. The *regret* of the sequence of measurements $E_1, \ldots, E_T$ is defined as

$$R_T := \sum_{t=1}^{T} \ell_t(\mathrm{Tr}(E_t\omega_t)) - \min_{\varphi \in \mathsf{D}(d)} \sum_{t=1}^{T} \ell_t(\mathrm{Tr}(E_t\varphi)).$$

Note that because the feedback provided by the adversary y_t need not be perfectly accurate, the loss function need not be consistent with any fixed quantum state. That is why we take the minimum over all mixed states $\varphi \in \mathsf{D}(d)$, rather than assuming this quantity is zero. We will first give an achievable bound on the regret after T iterations without proof and explain why this implies the desired mistake bound. Then for completeness we present an example of an algorithm which achieves the claimed bound on the regret after T iterations. The following two results are due to Ref. [4], with slight modification.

Theorem 2.1.3 (Regret bound)**.** *There exists an explicit rule for updating the estimates ω_t such that for any sequence of measurements $E_1, \ldots, E_T$, the regret is $R_T \leq 4\ln(d)/\epsilon + T\epsilon/4$.*

Corollary 2.1.4 (Mistake bound)**.** *Suppose the online learner applies the update rule from Theorem 2.1.3 whenever $\ell_t > 2\epsilon/3$, and outputs the previous estimate otherwise. Then the total number of mistakes is at most $M = O(\log(d)/\epsilon^2)$, independent of the number of rounds.*

Proof of Corollary 2.1.4. We have that $\sum_{t=1}^{T'} \ell_t(\mathrm{Tr}(E_t\omega_t)) \geq 2T'\epsilon/3$, where the sum is over the subsequence of iterations where an update is made. Similarly, the total loss of outputting the true state ρ in each of these iterations can be computed as $\sum_{t=1}^{T'} \ell_t(\mathrm{Tr}(E_t\rho)) \leq T'\epsilon/3$, since $\ell_t(\mathrm{Tr}(E_t\rho)) \leq \epsilon/3$ for any of these rounds t by assumption. Therefore $\min_{\varphi \in \mathsf{D}(d)} \sum_{t=1}^{T'} \ell_t(\mathrm{Tr}(E_t\varphi)) \leq T'\epsilon/3$. Together, these two bounds imply that the regret of the sequence of measurements $E_1, \ldots, E_{T'}$ corresponding to rounds t in which $\ell_t \geq 2\epsilon/3$ is $R_{T'} \geq T'\epsilon/3$. On the other hand, by Theorem 2.1.3 the total regret of this sequence of measurements is at most $R_{T'} \leq 4\ln(d)/\epsilon + T'\epsilon/4$. Combining with the lower bound on the regret we find $T' = O(\log(d)/\epsilon^2)$, and the result follows from the observation that the number of mistakes is $M \leq T'$. □

As mentioned previously, there are a few different algorithms which constitute proofs of Theorem 2.1.3. For completeness, we choose to explain how the MMW algorithm achieves the desired bound, summarizing Section 3 in [9] and Appendix C in [4].

Proof of Theorem 2.1.3

Consider the following matrix multiplicative weights (MMW) algorithm due to [8] and based on the earlier work of [35]. The algorithm takes as input a sequence of *loss matrices* L_t, and Theorem 2.1.3 will follow upon defining these appropriately based on the loss in the quantum online learning setting.

> **Matrix multiplicative weights (MMW) algorithm**. Given parameter $\eta \in (0,1)$ and a sequence of *loss matrices* $0 \preceq L_t \preceq \mathbb{1}$, in each round $t = 1, 2, \ldots$:
>
> 1. Compute $W_t := \exp(-\eta \sum_{\tau=1}^{t-1} L_\tau)$. (We have $W_1 = \mathbb{1}$.)
> 2. Output the estimate $\omega_t := \frac{W_t}{\mathrm{Tr}(W_t)}$.
> 3. Receive the loss matrix L_t.

Lemma 2.1.5 (Theorem 3.1 in [9])**.** *The MMW algorithm outputs estimates ω_t satisfying*

$$\sum_{t=1}^{T} \mathrm{Tr}(L_t \omega_t) \leq \lambda_d \left(\sum_{t=1}^{T} L_t \right) + \eta \sum_{t=1}^{T} \mathrm{Tr}(L_t^2 \omega_t) + \frac{\ln(d)}{\eta}$$

where $\lambda_d(\cdot)$ denotes the minimum eigenvalue of a Hermitian matrix.

Proof. Following the proof in Section 3 of Ref. [9] we track the changes in the quantity $\mathrm{Tr}(W_t)$ for $t = 1, \ldots, T$ (known as a potential function in the online learning literature [7]). We have for any t that

$$\begin{aligned} \mathrm{Tr}(W_{t+1}) &= \mathrm{Tr}\left(\exp\left(-\eta \sum_{\tau=1}^{t} L_\tau \right) \right) \\ &\leq \mathrm{Tr}\left(\exp\left(-\eta \sum_{\tau=1}^{t} L_\tau \right) \exp(-\eta L_t) \right) \\ &= \mathrm{Tr}(W_t \exp(-\eta L_t)) \end{aligned}$$

where the inequality follows from the Golden-Thompson inequality (see Appendix B.2) $\mathrm{Tr}(\exp(A+B)) \leq \mathrm{Tr}(\exp(A)\exp(B))$ for two Hermitian matrices A, B. Then by the matrix inequality $\exp(-A) \preceq \mathbb{1} - A + A^2$ for any $\|A\| \leq 1$ (see Appendix B.2) we find that the right-hand side of the above is at most

$$\mathrm{Tr}(W_t)\left(1 - \eta \mathrm{Tr}(L_t \omega_t) + \eta^2 \mathrm{Tr}(L_t^2 \omega_t)\right) \leq \mathrm{Tr}(W_t) \exp\left(\eta \mathrm{Tr}(L_t \omega_t) + \eta^2 \mathrm{Tr}(L_t^2 \omega_t)\right)$$

using the inequality $\mathrm{e}^x \geq 1 + x$ for all real x. By induction on t with base case $\mathrm{Tr}(W_1) = \mathrm{Tr}(\mathbb{1}) = d$ we arrive at

$$\mathrm{Tr}(W_{T+1}) \leq d \exp\left(-\eta \sum_{t=1}^{T} \mathrm{Tr}(L_t \omega_t) + \eta^2 \sum_{t=1}^{T} \mathrm{Tr}(L_t^2 \omega_t)\right)$$

while on the other hand, the fact that $\mathrm{Tr}(\exp(A)) = \sum_{k=1}^{d} \mathrm{e}^{\lambda_k(A)} \geq \mathrm{e}^{\lambda_d(A)}$ for any Hermitian matrix A implies that

$$\mathrm{Tr}(W_{T+1}) = \mathrm{Tr}\left(\exp\left(-\eta \sum_{t=1}^{T} L_t\right)\right) \geq \exp\left(-\eta \lambda_d \left(\sum_{t=1}^{T} L_t\right)\right).$$

The result follows upon combining the two inequalities and taking logarithms of both sides. □

We can now return to proving Theorem 2.1.3. Firstly, Lemma 2.1.5 implies that

$$\sum_{t=1}^{T} \mathrm{Tr}(L_t \omega_t) \leq \sum_{t=1}^{T} \mathrm{Tr}(L_t \zeta) + \eta T + \frac{\ln(d)}{\eta} \tag{2.1}$$

for any density matrix ζ, where we have used the fact that $\lambda_d(A) = \min_{\varphi \in \mathsf{D}(d)} \mathrm{Tr}(A\varphi)$ for any Hermitian positive semidefinite matrix A, along with the inequality $\mathrm{Tr}(L_t^2 \omega_t) \leq 1$. Define the function[1] $\ell_t'(x) = (2\mathbf{1}_{x \geq y_t} - 1)$, i.e., equal to $+1$ if $x \geq y_t$ and -1 otherwise. If $x \geq y_t$ then for any $z \in \mathbb{R}$ we have $\ell_t(x) - \ell_t(z) = x - y_t - |z - y_t| \leq x - z$. If $x < 0$ then $\ell_t(x) - \ell_t(z) = y_t - x - |z - y_t| \leq -x + z$. Using the definition of ℓ_t', these two inequalities can be expressed as $\ell_t(x) - \ell_t(z) \leq \ell_t'(x)(x - z)$ for any $x, z \in \mathbb{R}$. Let the loss matrices be defined as $L_t = \ell_t'(\mathrm{Tr}(E_t \omega_t))E_t$ for each iteration t. We may substitute into the inequality (2.1) to deduce that

$$\begin{aligned}
\eta T + \frac{\ln(d)}{\eta} &\geq \sum_{t=1}^{T} \ell_t'(\mathrm{Tr}(E_t \omega_t))\mathrm{Tr}(E_t \omega_t) - \sum_{t=1}^{T} \ell_t'(\mathrm{Tr}(E_t \omega_t))\mathrm{Tr}(E_t \zeta) \\
&\geq \sum_{t=1}^{T} \ell_t(\mathrm{Tr}(E_t \omega_t)) - \sum_{t=1}^{T} \ell_t(\mathrm{Tr}(E_t \zeta))
\end{aligned}$$

[1] This is the *subgradient* of the loss function, an important tool in convex optimization to extend the definition of gradients beyond differentiable functions. Since we are only dealing with the special case where $\ell_t(x) = |x - y_t|$, we do not elaborate on this point here, but refer the interested reader to Chapter D. in Ref. [21].

where the second line follows from the inequality $\ell_t(x) - \ell_t(z) \leq \ell_t'(x)(x - z)$ for any $x, z \in \mathbb{R}$. Choosing $\eta = \epsilon/4$ and noting this inequality holds for any density matrix ζ gives the desired result.

2.2 Tomography with binary measurements and an online learner

Recall from Section 1.5 that there is a straightforward method to perform tomography using $O(d^4/\epsilon^2)$ copies of the unknown state, by estimating the expectations of the various Pauli operators. We will see in Section 3.2.2 that this is in fact the optimal sample complexity using binary measurements under the nonadaptive measurement assumption. In this section, we show how to achieve a matching sample complexity in terms of d by letting the experimentalist act as the adversary who provides feedback to the online learner. Clearly, a sequence of uninformative measurements (e.g., $\mathbb{1}, \mathbb{1}, \dots$) prevents tomography of the state. However, the experimentalist performing the measurements may choose to be as informative as possible with their sequence, stepping out of their role as adversary and becoming more of a "teacher" in order to enable the tomography. Thus, the question we are concerned with is: what must the adversary do to guarantee that the learner's estimates are eventually close to the true state in some metric? The procedure described in Algorithm 1 along with Theorem 2.2.1 provide an answer.

Theorem 2.2.1. *Consider the procedure described in Algorithm 1. Let $\rho \in \mathsf{D}(d)$, $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, and M be the mistake bound of the online learner with mistake parameter ϵ, as in Corollary 2.1.4. Let p be the distribution of the random measurement $\boldsymbol{U}Q\boldsymbol{U}^\dagger$ where Q is a rank-$d/2$ orthogonal projection and $\boldsymbol{U}$ is a random unitary comprising a 4-design. For $T = M$ and some $K = O(\log(M/\delta))$ it holds that the output ω satisfies*

$$\frac{1}{\sqrt{d}} \left\| \omega - \rho \right\|_{\mathrm{F}} = O\left(\epsilon\right)$$

with probability at least $1 - \delta$.

Let us pause here to explain why this theorem implies that tomography with $\sim d^4$ copies of the state is possible. We give a more rigorous treatment at the end of this section. Line 8 in Algorithm 1 calls for $O(\epsilon)$-accurate estimates of the expectation $\mathrm{Tr}(E_i\rho)$, which can be accomplished by performing the binary measurement corresponding to $\{E_i, \mathbb{1} - E_i\}$ a number of times which is $O(1/\epsilon^2)$. Furthermore, we know from Corollary 2.1.4 that $M \approx \log(d)/\epsilon^2$ is achievable, so the sample complexity of achieving the condition in

Algorithm 1 Tomography from online learning

Input Unknown state ρ, mistake parameter $\epsilon \in (0,1)$, confidence $\delta \in (0,1)$, K, T, distribution over measurements p

Output Estimate ω

for $t = 1, \ldots, T$ **do**
  receive ω_t from online learner
  for $i = 1, \ldots, K$ **do** ▷ The helpful adversary
    if $i = K$ **then**
      return ω_t
    end if
    sample $E_i \sim p$
    $y_i \leftarrow$ estimate of $\mathrm{Tr}(E_i \rho)$ with $\epsilon/3$ accuracy
    if $|y_i - \mathrm{Tr}(E_i \omega_t)| > 4\epsilon/3$ **then**
      $E_t \leftarrow E_i, y_t \leftarrow y_i$
      break
    end if
  end for
  send measurement E_t and feedback y_t to online learner
end for
return ω_T

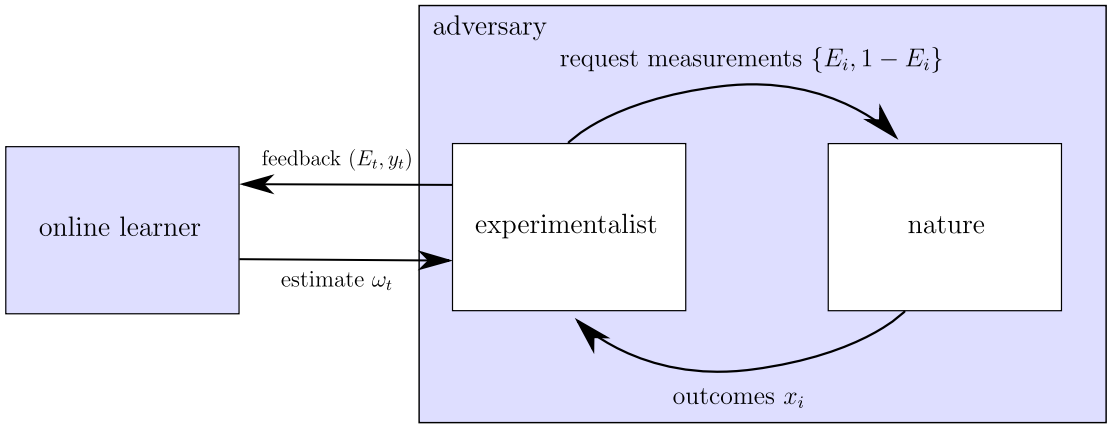

Figure 2.1: Schematic of online learning as a black-box for full state tomography.

Theorem 2.2.1 is (ignoring $\log\log(d)$ and $\log(1/\epsilon)$ factors) $N \approx \log(d)/\epsilon^4$. By the inequality $\|\cdot\|_1 \leq \sqrt{d}\,\|\cdot\|_{\mathrm{F}}$, making the replacement $\epsilon \to \epsilon/d$ leads to an estimate ω which is accurate to within ϵ in trace distance, for a total sample complexity of $\sim d^4/\epsilon^4$

We will now establish some helpful lemmas before proceeding with the proof of Theorem 2.2.1.

Lemma 2.2.2 (Paley-Zigmund inequality)**.** *For a real-valued random variable $\boldsymbol{Z}$ and fixed $\theta \in (0, 1)$, it holds that*

$$\Pr\left[|\boldsymbol{Z}| > \sqrt{\theta \mathbb{E}\boldsymbol{Z}^2}\right] \geq (1-\theta)^2 \frac{(\mathbb{E}\boldsymbol{Z}^2)^2}{\mathbb{E}\boldsymbol{Z}^4}.$$

Proof. For a nonnegative random variable $\boldsymbol{X}$ and parameter $\theta \in (0, 1)$, let E denote the event $\boldsymbol{X} \leq \theta\,\mathbb{E}\,\boldsymbol{X}$, $\overline{E}$ its complement, and let $\mathbf{1}_E$ denote the indicator random variable for the event E. We have

$$\mathbb{E}\,\boldsymbol{X} = \mathbb{E}\,\boldsymbol{X}(\mathbf{1}_E + \mathbf{1}_{\overline{E}}) \leq \theta\,\mathbb{E}\,\boldsymbol{X} + (\mathbb{E}\,\boldsymbol{X}^2)^{1/2}\Pr\left[\boldsymbol{X} > \theta\,\mathbb{E}\,\boldsymbol{X}\right]^{1/2}$$

where we have used Cauchy-Schwarz inequality to arrive at the second term in the inequality. Letting $\boldsymbol{X} = \boldsymbol{Z}^2$ and rearranging gives us the desired inequality. □

The following lemma combined with Lemma 2.2.2 allows us to lower bound the probability that a mistake occurs whenever ω_t is sufficiently far from the true state. The proof is provided in Appendix A and relies on the measurements being constructed from unitary 4-designs. Briefly, the reason for this is that 4-designs constitute a sufficiently expressive

set of measurements for distinguishing between two states that are far enough apart in the metric induced by the Frobenius norm. (See for example [5] for another application of this property.) In the language of online learning, large distances between the estimate and the true state manifest in sufficiently high probabilities of a mistake occurring.

Lemma 2.2.3. *Let Q be a fixed rank-$d/2$ orthogonal projection and consider a random measurement operator $\boldsymbol{E}$ where $\boldsymbol{E} := \boldsymbol{U}Q\boldsymbol{U}^\dagger$ for $\boldsymbol{U}$ a random unitary comprising a 4-design. Define $\boldsymbol{Z} = \mathrm{Tr}(\boldsymbol{E}(\omega - \rho))$ for some fixed density matrices ρ, ω. The following relations hold:*

$$\frac{c}{d}\left\|\omega-\rho\right\|_{\mathrm{F}}^2 \leq \mathbb{E}\boldsymbol{Z}^2 \leq \frac{C}{d}\left\|\omega-\rho\right\|_{\mathrm{F}}^2, \qquad \mathbb{E}\boldsymbol{Z}^4 \leq \frac{c'}{d^2}\left\|\omega-\rho\right\|_{\mathrm{F}}^4$$

where c, C, and c' are universal constants.

We are now ready to prove Theorem 2.2.1.

Proof of Theorem 2.2.1. We refer to the indices t of the first loop in Algorithm 1 as the "rounds". First note that each measurement E_t the online learner receives in round t is a measurement for which they are guaranteed to make a mistake. This is since each of these measurements must satisfy $|y_t - \mathrm{Tr}(E_t\omega_t)| > 4\epsilon/3$ by Line 9, which by triangle inequality along with the $\epsilon/3$ accuracy of y_t implies $|\mathrm{Tr}(E_t\rho) - \mathrm{Tr}(E_t\omega_t)| > \epsilon$.

Hence if the algorithm terminates and outputs ω_T this estimate will cease to make mistakes on any future measurements because $T = M$ and the mistake bound is M. In other words, if the algorithm returns ω_T we must have that $\frac{1}{2}\left\|\omega_T - \rho\right\|_1 = \max_{0\preceq E\preceq \mathbb{1}} \mathrm{Tr}(E(\omega_T - \rho)) \leq \epsilon$ which certainly implies the relation in Theorem 2.2.1. So it suffices to upper bound the probability that the relation does not hold when the algorithm returns ω_t for some $t < M$. Suppose $\sqrt{\frac{c}{2d}}\left\|\omega_t - \rho\right\|_{\mathrm{F}} > 5\epsilon/3$ and let $\boldsymbol{Z}_i := \mathrm{Tr}(\boldsymbol{E}_i(\omega_t - \rho))$ for $\boldsymbol{E}_i$ the random measurement in Line 7 of the algorithm, which is based on a 4-design. By applying Lemma 2.2.2 with $\theta = 1/2$ and combining with Lemma 2.2.3 we obtain

$$\Pr\left[|\boldsymbol{Z}_i| > 5\epsilon/3\right] \geq \Pr\left[|\boldsymbol{Z}_i| > \sqrt{\frac{c}{2d}}\left\|\omega_t - \rho\right\|_{\mathrm{F}}\right] \geq c'' \tag{2.2}$$

for some absolute constant c''. Also, we have that $\boldsymbol{Z}_i > 5\epsilon/3 \implies |y_i - \mathrm{Tr}(\boldsymbol{E}_i\omega_t)| > 4\epsilon/3$. Therefore, if the algorithm returns the state ω_t it is because in round t the random variables $\boldsymbol{Z}_i \leq 5\epsilon/3$ for every $i \in [K]$. The probability of this occurring is at most $(1-c'')^K \leq \mathrm{e}^{-c''K}$ by (2.2). Therefore, taking $K = \ln(M/\delta)/c''$ and applying union bound we find the probability is at most δ that in any round t the algorithm returns an estimate ω_t such that $\sqrt{\frac{c}{2d}}\left\|\omega_t - \rho\right\|_{\mathrm{F}} > 5\epsilon/3$. □

Further details on sample complexity upper bound

The remainder of this section explains in greater detail how Theorem 2.2.1 implies that tomography to within ϵ trace distance can be accomplished with $\tilde{O}(d^4/\epsilon^4)$ copies of the state ρ. Let us first consider which line of Algorithm 1 actually requires measurements of this state. Upon inspection, only Line 8 requires measurements of ρ. In particular, for a fixed measurement operator sampled in Line 7 E_i, it suffices to perform the measurement corresponding to $\{E_i, \mathbb{1} - E_i\}$ a number of times $N = \frac{1}{2(\epsilon/3)^2} \log(2/p)$ in order to estimate the expectation $\mathrm{Tr}(E_i \rho)$ to within $\epsilon/3$ with probability at least $1 - p$, using Hoeffding's inequality. Consider the quantities T and K which are inputs to the algorithm. The maximum number of times that Line 8 can be called is equal to TK. Therefore, by union bound, taking $p = \delta'/(TK)$ in the above argument implies that the probability that any call to Line 8 fails to produce an $\epsilon/3$-accurate estimate is at most δ'. Let us call this failure event A. From Theorem 2.2.1, we know that, given the event $\neg A$ occurs (no failures in the estimation step), the algorithm produces an estimate ω satisfying $\frac{1}{\sqrt{d}} \|\omega - \rho\|_{\mathrm{F}} = O(\epsilon)$ with failure probability at most δ'' so long as $T = M \leq O(\log(d)/\epsilon^2)$ (the mistake bound with mistake parameter ϵ) and $K = O(\log(M/\delta''))$. Here, we have assumed that the online learner uses the updates specified in Corollary 2.1.4. The total failure probability is thus at most $\Pr[A] + \Pr[B|\neg A] \leq \delta' + \delta''$. Taking $\delta' = \delta'' = \delta/2$ we find that using a total of at most

$$TKN = O\left(\frac{\log(d)}{\epsilon^4} \log\left(\frac{\log(d)}{\epsilon^2 \delta}\right)\right)$$

copies of the state suffices to achieve the Frobenius norm condition in Theorem 2.2.1 with probability at least $1 - \delta$. Then, making the substitution $\epsilon \to \epsilon/d$ and using the matrix inequality $\|\cdot\|_1 \leq \sqrt{d} \|\cdot\|_{\mathrm{F}}$ gives the desired upper bound on the sample complexity of performing tomography to ϵ-accuracy in trace distance in this manner.

Chapter 3

Lower bounds with nonadaptive measurements

In this chapter we show lower bounds for quantum state tomography in the nonadaptive measurement model which we introduced in Section 1.4. In the interest of worst-case bounds, we do not make any assumptions regarding the rank of the state to be learned although in theory knowledge of the rank can lead to improvements in the sample complexity [16].

We begin with a $\Omega(d^2/\epsilon^2)$ lower bound for the entangled measurement setting, in which we describe the setup employed for the remaining sections. This flavour of lower bound has been known since at least 2015 [31], although the basic argument in the context of tomography goes back further [16]. Having since been employed in a variety of different contexts, the basic argument involves constructing a sufficiently large ϵ-packing of states and then using Fano's inequality to bound the mutual information between the measurement statistics and the choice of state from the packing. An often neglected fact in the quantum information literature is that this technique is standard in the sub-field of *density estimation* – which is the classical analogue of quantum tomography – going back to at least 1978 [25].

Following the setup, we give a proof of a result due to Haah et al. [18] that $\Omega(d^3/\epsilon^2)$ copies of the state are required with nonadaptive measurements. For the case of nonadaptive constant-outcome measurements, we derive a lower bound of $\Omega(d^4/\epsilon^2)$ copies which to the best of our knowledge has not appeared in the literature. This implies that the straightforward tomography scheme using binary Pauli measurements outlined in Section 1.5 is optimal in this setting.

3.1 Entangled measurements and Holevo's theorem

In this section, we will prove the following theorem, which is implied by Lemma 5 in Haah et al. [18].

Theorem 3.1.1. *Suppose a tomography algorithm has the following behaviour for any mixed state $\rho \in \mathsf{D}(d)$: given a register in the state $\rho^{\otimes n}$, the algorithm outputs an estimate $\hat{\rho}$ such that $\|\hat{\rho} - \rho\|_1 \leq \epsilon$ with at least a constant probability of success. Then, it must hold that the algorithm uses $n = \Omega\left(d^2/\epsilon^2\right)$ samples of the state.*

To demonstrate lower bounds for quantum tomography, it suffices to show that there exists a large, but well-separated collection (an ϵ-packing) of quantum states which is 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. Our approach throughout this section will therefore be to construct a hard instance of the state discrimination problem, and then argue that if the number of copies n is too small the success probability of our protocol goes to zero as the parameters d and $1/\epsilon$ increase.

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

Our ϵ-packing will be comprised of states of the following form (also considered in the lower bounds of [18, 13]):

$$\rho_{\epsilon,U} := \epsilon U \sigma U^\dagger + (1-\epsilon)\frac{\mathbb{1}}{d} \tag{3.1}$$

where $\sigma := \frac{2}{d}Q$ for Q, a fixed rank-$d/2$ orthogonal projection and $\epsilon \in (0,1)$ is some parameter interpolating between σ and the completely mixed state. (Assuming d is even

here does not take away from the argument, and we may proceed analogously with a floor or ceiling.) Intuitively, such states are useful for proving lower bounds for the task of learning an unknown state because they represent a hard case where the completely mixed state is slightly perturbed, which leads to noisy measurement statistics. We will make use of the definition in Eq. (3.1) often in the remainder of this thesis.

How does one construct an ϵ-packing of states of this form? We can 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 apply the argument many times recursively. The concentration inequalities we will invoke to arrive at our tail bounds follow from log-Sobolev inequalities, analogous to Lévy's Lemma for functions on the unit sphere [28]. 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 lemma, which we do not prove.

Lemma 3.1.2 (Consequence of Theorems 5.5 and 5.16 from [28])**.** *Let $f : \mathbb{U}(d) \to \mathbb{R}$ be an L-Lipschitz function 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*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[|f(\boldsymbol{U}) - \mu| \geq t\right] \leq 2\exp\left(-\frac{dt^2}{12L^2}\right).$$

This concentration inequality enables us to derive a variation of a result which is due to Ref. [19], called "concentration of projector overlaps". We will invoke this lemma repeatedly in the remainder of this thesis in order to claim the existence of various collections of quantum states.

Lemma 3.1.3 (Concentration of projector overlaps)**.** *Let $\boldsymbol{U}$ be a Haar-random unitary operator taking values in $\mathbb{U}(d)$ and let $P, Q \in \mathsf{Psd}(d)$ be orthogonal projective operators with rank r_P,r_Q respectively. It holds that*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\left|\mathrm{Tr}(P\boldsymbol{U}Q\boldsymbol{U}^\dagger) - \frac{r_P r_Q}{d}\right| \geq t\right] \leq 2\exp\left(-\frac{cdt^2}{\sqrt{r_P r_Q}}\right)$$

where c is some universal constant.

Proof. Define $f : \mathbb{U}(d) \to \mathbb{R}$ by $f(U) = \mathrm{Tr}(PUQU^\dagger)$ for all $U \in \mathbb{U}(d)$. We will show that the expectation of this function is $r_P r_Q/d$ and that it is $O((r_P r_Q)^{1/4})$-Lipschitz.

For the expectation, we have from the first relation in Proposition A.0.2 that $\mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} \boldsymbol{U}Q\boldsymbol{U}^\dagger = r_Q\mathbb{1}/d$. Then by linearity of trace we have $\mathbb{E}\,\mathrm{Tr}(P\boldsymbol{U}Q\boldsymbol{U}^\dagger) = \mathrm{Tr}(P)r_Q/d = r_P r_Q/d$.

For the Lipschitz constant, consider the difference $f(U)-f(V)$ for two arbitrary unitary operators $U, V \in \mathbb{U}(d)$. We have

$$\begin{aligned}
|f(U) - f(V)| &= \left|\mathrm{Tr}(P(UQU^\dagger - VQV^\dagger))\right| \\
&= \frac{1}{2}\left|\mathrm{Tr}(P(U+V)Q(U-V)^\dagger) + \mathrm{Tr}(P(U-V)Q(U+V)^\dagger)\right| \\
&\leq \frac{1}{2}\left|\mathrm{Tr}(P(U+V)Q(U-V)^\dagger)\right| + \frac{1}{2}\left|\mathrm{Tr}(P(U-V)Q(U+V)^\dagger)\right|. \qquad (3.2)
\end{aligned}$$

Let us focus on just the first term in (3.2). Neglecting the factor of $1/2$ we have

$$\begin{aligned}
|\mathrm{Tr}(P(U+V)Q(U-V)^\dagger)| &\leq |\mathrm{Tr}(PUQ(U-V)^\dagger)| + |\mathrm{Tr}(PVQ(U-V)^\dagger)| \\
&\leq (\|PUQ\|_{\mathrm{F}} + \|PVQ\|_{\mathrm{F}})\,\|U-V\|_{\mathrm{F}}
\end{aligned}$$

where we have used Cauchy-Schwarz to arrive at the second inequality. For any unitary operator W we have $\|PWQ\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(PWQW^\dagger)} \leq \sqrt{\|P\|_{\mathrm{F}}\,\|Q\|_{\mathrm{F}}} = \sqrt{\sqrt{r_P r_Q}}$ using cyclic property of trace along with the fact that P and Q are orthogonal projectors. This implies that the first term in (3.2) is at most $(r_P r_Q)^{1/4}\,\|U-V\|_{\mathrm{F}}$. A similar argument can be made for the second term, which gives the desired upper bound on the Lipschitz constant of f. □

We now give the lemma that we can use to construct a sufficiently large packing of quantum states of the form (3.1) which is difficult to discriminate. In conjunction with Holevo's theorem, this hard instance necessitates the $\Omega(d^2/\epsilon^2)$ lower bound. This is a special case of the approach that is adopted in Haah et al. [18].

Lemma 3.1.4. *There exists a universal constant c such that the following holds. Pick $\epsilon \in (0,1)$, let $d > 0$ be a positive integer, and let $0 \leq N < \mathrm{e}^{cd^2}/2$ be an integer. Consider a set of states $\{\rho_1, \rho_2, \ldots, \rho_N\} \subset \mathsf{D}(d)$ where*

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

for each $i \in [N]$, $U_1, U_2, \ldots, U_N \in \mathbb{U}(d)$ are arbitrary unitary operators, and σ is as in (3.1). For Haar-random $\boldsymbol{U}$ taking values in $\mathbb{U}(d)$, the probability that $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$ for any $i \in [N]$ is strictly less than 1.

Proof. Let $\boldsymbol{U}$ be a Haar-random unitary taking values in $\mathbb{U}(d)$, $Q \in \mathsf{Psd}(d)$ be the fixed orthogonal projection appearing in the definition of $\rho_{\epsilon,U}$ for arbitrary unitary operator U, and $P := \mathbb{1} - Q$. A straightforward consequence of Lemma 3.1.3, the concentration of projector overlaps lemma, is the following upper bound:

$$\Pr\left[\mathrm{Tr}(P\boldsymbol{U}Q\boldsymbol{U}^\dagger) \leq d/8\right] \leq 2\mathrm{e}^{-cd^2}$$

for some absolute constant c. This follows by taking $t = d/8$ in the lemma and combining with the fact that P, Q are both rank $d/2$. Also, using the definition of $\rho_{\epsilon,U}$ we have

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

for any $U \in \mathbb{U}(d)$. Now, due to the inequality $\|A\|_1 \geq |\mathrm{Tr}(AV)|$ for any $V \in \mathbb{U}(d)$ and $A \in \mathsf{H}(d)$, one may right-multiply the right-hand side in the above equation by the unitary $P - Q$ and take the trace to obtain

$$\begin{aligned}
\left\|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{1}}\right\|_1 &\geq \frac{2\epsilon}{d}\left|\mathrm{Tr}(UQU^\dagger P - UQU^\dagger Q + Q)\right| \\
&= \frac{2\epsilon}{d}\left|\mathrm{Tr}(UQU^\dagger P) + \mathrm{Tr}(UQU^\dagger(\mathbb{1} - Q))\right| \\
&= \frac{4\epsilon}{d}\mathrm{Tr}(PUQU^\dagger)
\end{aligned}$$

where in going from the first to the second line we have used linearity along with cyclic property of trace and the last line follows from the definition of P. Therefore, if $\left\|\rho_{\epsilon,U} - \rho_{\epsilon,\mathbb{1}}\right\|_1 \leq \epsilon/2$ for a unitary $U \in \mathbb{U}(d)$, we also have $\mathrm{Tr}(PUQU^\dagger) \leq d/8$, from which we may conclude that for a Haar-random unitary $\boldsymbol{U}$,

$$\Pr\left[\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{1}}\right\|_1 \leq \epsilon/2\right] \leq 2\mathrm{e}^{-cd^2}.$$

Next, consider the unitary U_i and corresponding state ρ_i in the lemma, for some $i \in [N]$. Using the invariance of the trace distance under unitary transformation, one may verify that for the Haar-random unitary $\boldsymbol{U}$

$$\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_{\epsilon,\mathbb{1}}\right\|_1 = \left\|\rho_{\epsilon,U_i\boldsymbol{U}} - \rho_i\right\|_1$$

which, by left-invariance of the Haar measure, leads to the conclusion that

$$\Pr\left[\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\right\|_1 \leq \epsilon/2\right] \leq 2\mathrm{e}^{-cd^2}. \tag{3.3}$$

Since this inequality holds for any index $i \in [N]$ the proof is complete upon applying union bound over the events $\left\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\right\|_1 \leq \epsilon/2$ for each $i \in [N]$. □

Lemma 3.1.4 implies we may construct a (non-explicit) ensemble of states of size $N = \exp(\Omega(d^2))$ which is an $\epsilon/2$-packing in trace distance, using a probabilistic existence argument.

Definition 3.1.5 (ϵ-packing condition)**.** A set of mixed states $\mathcal{S}$ satisfies the *ϵ-packing condition* for some $\epsilon > 0$ if it holds that $\|\rho - \rho'\|_1 > \epsilon$ for every $\rho, \rho' \in \mathcal{S}$ such that $\rho \neq \rho'$.

Corollary 3.1.6. *Let $d > 0$ be a positive integer. There exists a set of $N \geq \exp(\Omega(d^2))$ quantum states as in Lemma 3.1.4 for which the $\epsilon/2$-packing condition is satisfied.*

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 Lemma 3.1.4, where $k < \mathrm{e}^{cd^2 - \ln(2)}$ and c is as in Lemma 3.1.4. Suppose further that this set satisfies the packing condition; namely, $\|\rho_i - \rho_j\|_1 > \epsilon/2$ for all $i, j \in [k], i \neq j$. From Lemma 3.1.4 we know that the probability of choosing a unitary $\boldsymbol{U}$ Haar randomly such that $\mathcal{S}_k \cup \{\rho_{\epsilon,\boldsymbol{U}}\}$ no longer satisfies the $\epsilon/2$-packing condition is strictly less than one. Therefore, there must exist at least one state which we can add to the packing. The result follows by induction on k. □

Let us now turn our attention to bounding the Holevo information arising from an ensemble of states constructed using the corollary above.

Lemma 3.1.7. *Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subseteq \mathsf{D}(d)$ be a set of $N = \exp(\Omega(d^2))$ quantum states as in Corollary 3.1.6. Let $\boldsymbol{x}$ be uniformly random over $[N]$ and consider the ensemble of states corresponding to $\rho_{\boldsymbol{x}}^{\otimes n}$. The Holevo information for this ensemble satisfies*

$$S\left(\frac{1}{N}\sum_{i=1}^{N}\rho_i^{\otimes n}\right) - \frac{1}{N}\sum_{i=1}^{N} S\left(\rho_i^{\otimes n}\right) \leq n\epsilon^2.$$

Proof. Let $\rho := \frac{1}{N}\sum_{i=1}^{N}\rho_i^{\otimes n}$ be the ensemble state and note that the reduced state on any of the n registers is just $\tau := \frac{1}{N}\sum_{i=1}^{N}\rho_i$. It follows from the subadditivity of von Neumann entropy that $S(\rho) \leq nH(\tau)$. We also have $S\left(\rho_i^{\otimes n}\right) = nS(\rho_i)$, so it suffices to show the upper bound

$$S(\tau) - \frac{1}{N}\sum_{i=1}^{N} S(\rho_i) \leq \epsilon^2. \tag{3.4}$$

The first term on the left-hand side is at most $\log(d)$. Also, since each state ρ_i is of the form $\rho_i = \frac{2\epsilon}{d}U_i Q U_i^\dagger + (1-\epsilon)\frac{\mathbb{1}}{d}$ for some unitary U_i, half its eigenvalues are equal to $(1+\epsilon)/d$

and the other half are $(1-\epsilon)/d$. (This follows from the fact that Q is a rank-$d/2$ orthogonal projector.) Therefore, we may compute

$$S(\rho_i) = H((1+\epsilon)/2) + \log(d/2)$$

where the entropy on the right-hand side is the binary entropy function. Thus, the expression on the left-hand side of (3.4) is at most

$$\log(d) - \log(d/2) - H((1+\epsilon)/2) \leq 1 - (1-\epsilon^2) = \epsilon^2$$

where the first relation follows from the inequality for the binary entropy function $H(p) \geq 1 - 4\delta^2$ whenever $|p - 1/2| \leq \delta$. □

We can now prove Theorem 3.1.1. Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\}$ be an $\epsilon/2$-packing as described in the probabilistic existence argument of Corollary 3.1.6. Also, let $\boldsymbol{x}$ be uniform over $[N]$ and $\boldsymbol{y}$ be the outcome of a measurement on the state $\rho_{\boldsymbol{x}}^{\otimes n}$. By Holevo's theorem and Lemma 3.1.7, the mutual information between these random variables is upper bounded as

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq n\epsilon^2.$$

Assume that the measurement whose outcome is $\boldsymbol{y}$ succeeds in identifying the state $\rho_{\boldsymbol{x}}$ to within $\epsilon/2$ in trace distance with probability $\Omega(1)$, implying that we can decode $\boldsymbol{x}$ with at least a constant probability of success. Then, by Fano's inequality (Corollary 1.3.6), for sufficiently large d we must have

$$c_1 d^2 \leq c_2 \log(N) \leq I(\boldsymbol{x} : \boldsymbol{y}) \leq n\epsilon^2$$

for some universal constants c_1, c_2, which can only be true if $n = \Omega(d^2/\epsilon^2)$.

3.2 Nonadaptive measurements

We consider next the first of the unentangled measurement settings, where measurements are performed on individual copies of the state nonadaptively i.e., the k^{th} measurement does not depend on the previous $k-1$ outcomes. This means that the random variable $\boldsymbol{y}_k$ corresponding to the k^{th} measurement outcome is independent of all other outcomes, given the state being measured. In this case, one may prove stronger bounds on the mutual information than the one due to Holevo's theorem.

To this end, let us begin by observing that, due to the conditional independence of the measurement outcomes $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the mutual information is subadditive. Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ be a set of $N = \exp(\Omega(d^2))$ states as in Corollary 3.1.6. Let $\boldsymbol{x}$ be uniformly random over $[N]$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the outcomes obtained from measuring $\rho_{\boldsymbol{x}}^{\otimes n}$ with n nonadaptive measurements. We have

$$I(\boldsymbol{x} : \boldsymbol{y}) \leq \sum_{k=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_k). \tag{3.5}$$

Then, the following lemma enables us to bound each of the mutual information terms appearing in the sum by an expectation over $\boldsymbol{x}$.

Lemma 3.2.1. *Let $\boldsymbol{x}$ be an arbitrary random variable with marginal distribution $p_{\boldsymbol{x}}$ and $\boldsymbol{y}$ be a discrete random variable. Denote by $p_{\boldsymbol{y}|x}$ the distribution of $\boldsymbol{y}$ given a fixed value x of $\boldsymbol{x}$. For an arbitrary discrete distribution q defined on the same space as the distribution of $\boldsymbol{y}$, it holds that*

$$\ln(2) \times I(\boldsymbol{x} : \boldsymbol{y}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} \chi^2(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel q). \tag{3.6}$$

Proof. By Definition 1.3.1 we have $I(\boldsymbol{x} : \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel p_{\boldsymbol{y}})$ where $p_{\boldsymbol{y}}$ is the marginal distribution of $\boldsymbol{y}$. We also have the inequality $D_{\mathrm{KL}}(a \parallel b) \leq \log(e)\chi^2(a \parallel b)$ for any two discrete distributions a and b defined on the same space, as in Lemma 1.3.8. This implies the relation in (3.6) upon showing that

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel q) \geq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel p_{\boldsymbol{y}}). \tag{3.7}$$

Let $\mathcal{Y}$ be the set of values $\boldsymbol{y}$ can take. Using the definition of KL-divergence, we have

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

which proves the relation in (3.7), since $D_{\mathrm{KL}}(p_{\boldsymbol{y}} \parallel q) \geq 0$ and $H(\boldsymbol{y}) - \mathop{\mathbb{E}}_{\boldsymbol{x}' \sim p_{\boldsymbol{x}}} H(\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{x}') = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel p_{\boldsymbol{y}})$. □

Corollary 3.2.2. *Define $\boldsymbol{x}, \boldsymbol{y}, p_{\boldsymbol{x}}, p_{\boldsymbol{y}|x}$ as in Lemma 3.2.1 and let $p_{\boldsymbol{y}}$ be the marginal distribution of $\boldsymbol{y}$. It holds that*

$$\ln(2) \times I(\boldsymbol{x} : \boldsymbol{y}) \leq \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} \chi^2(p_{\boldsymbol{y}|\boldsymbol{x}} \parallel p_{\boldsymbol{y}}) = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{x}}} \mathop{\mathbb{E}}_{\boldsymbol{y}' \sim p_{\boldsymbol{y}}} \frac{p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}')^2}{p_{\boldsymbol{y}}(\boldsymbol{y}')^2} - 1. \tag{3.8}$$

Proof. Since Lemma 3.2.1 holds in particular when $q = p_{\boldsymbol{y}}$, we immediately obtain the inequality in (3.8). The second relation can be seen upon substituting the definition of the chi-squared divergence (Definition 1.3.7) between $p_{\boldsymbol{y}|x}$ and $p_{\boldsymbol{y}}$ for some fixed value x of $\boldsymbol{x}$. We have

$$\chi^2(p_{\boldsymbol{y}|x} \parallel p_{\boldsymbol{y}}) = \mathop{\mathbb{E}}_{\boldsymbol{y}' \sim p_{\boldsymbol{y}}} \left(\frac{p_{\boldsymbol{y}|x}(\boldsymbol{y}')}{p_{\boldsymbol{y}}(\boldsymbol{y}')} \right)^2 - 1$$

which completes the proof of the corollary, by taking the expectation over $\boldsymbol{x}$ drawn from the marginal distribution $p_{\boldsymbol{x}}$. □

Although these results could be applied directly to the mutual information terms in the sum in (3.5), it would be intractable to compute the expectation over $\boldsymbol{x}$ given that we do not explicitly know the states which comprise our ensemble. 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 result is also stated in the proof of Lemma 10 in Haah et al. [18].)

Proposition 3.2.3. *Let $\boldsymbol{U}$ be a Haar-random unitary taking values in $\mathbb{U}(d)$ and $\boldsymbol{z}$ be the outcome obtained upon measuring $\rho_{\epsilon,\boldsymbol{U}}^{\otimes n}$ with measurement $\mathcal{M}$, where $\rho_{\epsilon,U}$ is defined as in (3.1) for any $U \in \mathbb{U}(d)$. Let $N = \exp(\Omega(d^2))$ be a positive integer and $\boldsymbol{x}$ be uniformly random over $[N]$. There exists a set $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ of states of the form in Lemma 3.1.4 satisfying the $\epsilon/2$-packing condition and for which*

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

where $\boldsymbol{y}$ is the outcome obtained from measuring $\rho_{\boldsymbol{x}}^{\otimes n}$ with $\mathcal{M}$.

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

Proof. Consider a fixed set $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ of states of the form in Lemma 3.1.4 satisfying the $\epsilon/2$-packing condition, (which we know exists from Corollary 3.1.6) and 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 the $\epsilon/2$-packing condition on our set of quantum states continues to hold if we replace our set of unitary operators $\mathcal{U}$ with the set $W\mathcal{U}$ for an arbitrary unitary operator $W \in \mathbb{U}(d)$, which can be seen by unitary invariance of the trace norm and the fact that $\|\rho_i - \rho_j\|_1 = \epsilon\|U_i\sigma U_i^\dagger - U_j\sigma U_j^\dagger\|_1$. For any fixed unitary W, define $\boldsymbol{y}_W$ to be the outcome obtained by measuring $\rho_{\epsilon,WU_{\boldsymbol{x}}}^{\otimes n}$ with $\mathcal{M}$ (equivalent to choosing a unitary from the shifted ensemble uniformly at random and measuring the state which corresponds to it). We claim that

$$\mathop{\mathbb{E}}_{\boldsymbol{W}\sim\text{Haar}} I(\boldsymbol{x} : \boldsymbol{y}_{\boldsymbol{W}}) \leq I(\boldsymbol{U} : \boldsymbol{z}). \tag{3.9}$$

First, define $p_{\boldsymbol{y}|W,x}$ to be the distribution of $\boldsymbol{y}_W$ given $\boldsymbol{x} = x$. By the definition of mutual information, we have

$$I(\boldsymbol{x} : \boldsymbol{y}_W) = H(\mathop{\mathbb{E}}_{\boldsymbol{x}} p_{\boldsymbol{y}|W,\boldsymbol{x}}) - \mathop{\mathbb{E}}_{\boldsymbol{x}} H(p_{\boldsymbol{y}|W,\boldsymbol{x}}).$$

Using concavity of entropy along with the independence of $\boldsymbol{x}$ and choice of unitary $\boldsymbol{W}$, the left-hand side of (3.9) is then at most

$$H\left(\mathop{\mathbb{E}}_{\boldsymbol{x}}\mathop{\mathbb{E}}_{\boldsymbol{W}\sim\text{Haar}} p_{\boldsymbol{y}|\boldsymbol{W},\boldsymbol{x}}\right) - \mathop{\mathbb{E}}_{\boldsymbol{x}}\mathop{\mathbb{E}}_{\boldsymbol{W}\sim\text{Haar}} H(p_{\boldsymbol{y}|\boldsymbol{W},\boldsymbol{x}}).$$

By right-invariance of the Haar measure, the expectation of some function of Haar-random $\boldsymbol{W}$ is equal to that expectation when $\boldsymbol{W}$ is mapped to $\boldsymbol{W}U^\dagger$, for any unitary operator U. Therefore, for any $x \in [N]$ we have

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

where $p_{\boldsymbol{z}|W}$ is the distribution of $\boldsymbol{z}$ given $\boldsymbol{U} = W$ and $p_{\boldsymbol{z}}$ is the marginal distribution of $\boldsymbol{z}$. Similarly, we have for any $x \in [N]$ that

$$\mathop{\mathbb{E}}_{\boldsymbol{W}\sim\text{Haar}} H(p_{\boldsymbol{y}|\boldsymbol{W},x}) = \mathop{\mathbb{E}}_{\boldsymbol{W}\sim\text{Haar}} H(p_{\boldsymbol{z}|\boldsymbol{W}}).$$

By the definition of mutual information, this proves the inequality (3.9). We may then 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 $I(\boldsymbol{U} : \boldsymbol{z})$, there must exist at least one unitary $V \in \mathbb{U}(d)$ for which the inequality $I(\boldsymbol{x} : \boldsymbol{y}_V) \leq I(\boldsymbol{U} : \boldsymbol{z})$ holds. The proposition follows by taking the set of density operators in the statement of the proposition to be $\mathcal{S}' = \{\rho_{\epsilon,VU_1}, \rho_{\epsilon,VU_2}, \ldots, \rho_{\epsilon,VU_N}\}$. □

We now move on to presenting two lower bounds for the nonadaptive case.

3.2.1 Arbitrary nonadaptive POVMs

In this section, we prove a lower bound in the nonadaptive case that is originally due to Ref. [18], where the POVMs acting on each state may be arbitrary.

Theorem 3.2.4 (Theorem 4 in Ref. [18])**.** *Consider a tomography algorithm in the nonadaptive measurement model which outputs an estimate $\hat{\rho} \in \mathsf{D}(d)$ such that $\|\hat{\rho} - \rho\|_1 \leq \epsilon$ with at least a constant probability of success, for any unknown state $\rho \in \mathsf{D}(d)$. It must hold that the algorithm uses $n = \Omega\left(d^3/\epsilon^2\right)$ samples of the state.*

Our analysis is simplified due to Lemma 3.2.1 as well as techniques for Haar integration based on permutation invariance. (We refer the interested reader to Section 7.2 of Ref. [38] for more on this topic.) We also do not assume the measurement operators are rank-one. This allows us to conclude our novel $\Omega(d^4/\epsilon^2)$ lower bound in the next section, in addition to laying the groundwork for what is to appear in Section 4.2.

Proof of Theorem 3.2.4

Another way of phrasing the above result is that it takes $\Omega(d^3/\epsilon^2)$ copies of a d-dimensional quantum state to learn it to within trace distance ϵ by using product measurements which are decided upon beforehand. We may write $\mathcal{M} = \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$ for some measurements $\mathcal{M}_1, \ldots, \mathcal{M}_n$ each of which acts on operators in $\mathsf{D}(d)$. Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ be the set of $N = \exp(\Omega(d^2))$ states which satisfy the $\epsilon/2$-packing condition as well as the mutual information claim in the statement of Proposition 3.2.3. By Fano's Inequality we know that the mutual information must satisfy $I(\boldsymbol{x} : \boldsymbol{y}) \geq \Omega(d^2)$ in order to discriminate the states with constant success probability, where $\boldsymbol{x}$ is uniform over $[N]$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ are the outcomes from performing the measurement $\mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$ on $\rho_{\boldsymbol{x}}^{\otimes n}$.

On the other hand, by Proposition 3.2.3 we have that $I(\boldsymbol{x} : \boldsymbol{y}) \leq I(\boldsymbol{U} : \boldsymbol{z})$ where $\boldsymbol{U}$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are defined as in the proposition: $\boldsymbol{U}$ is a Haar-random unitary, and $\boldsymbol{z}_i$ is the measurement outcome obtained by measuring $\rho_{\epsilon,\boldsymbol{U}}$ with $\mathcal{M}_i$, for each $i \in [n]$. Thus, it suffices to show $I(\boldsymbol{U} : \boldsymbol{z}) = o(d^2)$ for $n = o(d^3/\epsilon^2)$. Furthermore, by subadditivity of mutual information it holds that $I(\boldsymbol{U} : \boldsymbol{z}) \leq \sum_{k=1}^{n} I(\boldsymbol{U} : \boldsymbol{z}_k)$ so that by applying Corollary 3.2.2 (which is our χ^2-divergence upper bound on the mutual information) to each of the terms we arrive at

$$I(\boldsymbol{U} : \boldsymbol{z}_k) \leq \log(e) \left[\mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{\boldsymbol{z}_k}} \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \frac{p_{\boldsymbol{z}_k|\boldsymbol{U}}(\boldsymbol{z})^2}{p_{\boldsymbol{z}_k}(\boldsymbol{z})^2} - 1 \right] \tag{3.10}$$

for all $k \in [n]$ where for fixed $U \in \mathbb{U}(d)$ the conditional probabilities may be written as $p_{\boldsymbol{z}_k|U}(z) = \mathrm{Tr}(M_z^{(k)}\rho_{\epsilon,U})$ for some set of measurement operators $\{M_z^{(k)}\}_z$ corresponding to the measurement $\mathcal{M}_k$ on the k^{th} copy of the state. The marginal probabilities in the denominator are $p_{\boldsymbol{z}_k}(z) = \mathbb{E}_{\boldsymbol{U}\sim\text{Haar}}\mathrm{Tr}(M_z^{(k)}\rho_{\epsilon,\boldsymbol{U}})$ for each outcome z.

It then suffices to show that the Haar expectation in (3.10) at most $1+O(\epsilon^2/d)$ regardless of the measurement $\mathcal{M}_k$, for then we have that $I(\boldsymbol{U} : \boldsymbol{z}) \leq O(n\epsilon^2/d)$ which fulfills the requirement of being $o(d^2)$ when $n = o(d^3/\epsilon^2)$. In summary, we must show the following.

Proposition 3.2.5. *Let $\boldsymbol{z}$ be the random variable corresponding to the outcome of a measurement $\mathcal{M}$ performed on the state $\rho_{\epsilon,\boldsymbol{U}}$, where $\boldsymbol{U}$ is Haar-random and $\rho_{\epsilon,U}$ is as in (3.1), for each $U \in \mathbb{U}(d)$. It holds that*

$$\ln(2) \times I(\boldsymbol{U} : \boldsymbol{z}) \leq \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\text{Haar}} \chi^2(p_{\boldsymbol{z}|\boldsymbol{U}} \parallel p_{\boldsymbol{z}}) = O\left(\frac{\epsilon^2}{d}\right).$$

This proposition follows immediately from the next lemma.

Lemma 3.2.6. *Let $d > 1$ be a positive integer, $M \in \mathsf{Psd}(d)$, $0 \preceq M \preceq 1$ be a measurement operator, and $\rho_{\epsilon,U}$ be defined as in (3.1), for each $U \in \mathbb{U}(d)$ and $\epsilon \in (0,1)$. Also, let $w = \mathrm{Tr}(M)/d$. It holds that*

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

and

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

Proof. We will defer the calculation of some Haar integrals to Appendix A. By the definition of $\rho_{\epsilon,U}$ in (3.1) the first expectation is

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

Also, from the first Haar integral in Proposition A.0.2 in Appendix A along with linearity of trace we have

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

which leads to the first identity in the lemma. (Recall that Q is a rank-$d/2$ orthogonal projector.) For the second expectation in the lemma, note that by substituting the definition of $\rho_{\epsilon,U}$ and expanding we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}(M\rho_{\epsilon,\boldsymbol{U}})^2 = \frac{4\epsilon^2}{d^2} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}(M\boldsymbol{U}Q\boldsymbol{U}^\dagger)^2 + w^2(1-\epsilon^2). \tag{3.11}$$

Also, by linearity of trace and the fact that $\mathrm{Tr}(A)^2 = \mathrm{Tr}(A^{\otimes 2})$ for any $A \in \mathsf{H}(d)$ we have

$$\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{Tr}(M\boldsymbol{U}Q\boldsymbol{U}^\dagger)^2 = \mathrm{Tr}\left(M^{\otimes 2} \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} (\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes 2}\right). \tag{3.12}$$

The Haar integral on the right-hand side is evaluated explicitly in Proposition A.0.2 by setting the rank parameter r in that proposition to $d/2$,

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

where the identity and swap operation W in the above act on the space $(\mathbb{C}^d)^{\otimes 2}$. Substituting into (3.12) and making use of the identity $\mathrm{Tr}(W(A\otimes B)) = \mathrm{Tr}(AB)$ we find that the expectation value in (3.11) is equal to

$$\begin{aligned} \frac{\epsilon^2 w^2(d^2-2)}{d^2-1} + \frac{\epsilon^2 \mathrm{Tr}(M^2)}{d(d^2-1)} + w^2(1-\epsilon^2) &\leq \frac{\epsilon^2 w^2(d^2-2)}{d^2-1} + \frac{\epsilon^2 d w^2}{d^2-1} + w^2(1-\epsilon^2) \\ &= w^2\left(1 + \frac{\epsilon^2}{d+1}\right) \end{aligned} \tag{3.13}$$

as required, where the inequality follows from the fact that $\|M\|_1 \geq \|M\|_{\mathrm{F}}$ and M is Hermitian positive semidefinite. □

By substituting the two relations in Lemma 3.2.6 into the upper bound on the mutual information in (3.10) we find that $I(\boldsymbol{U} : \boldsymbol{Z}_k) \leq O(\epsilon^2/d)$, which implies result in Theorem 3.2.4.

3.2.2 Constant-outcome case

We can derive a stronger lower bound on the number of copies required in the nonadaptive setting when the measurements are restricted to having a constant number of outcomes. As a consequence we find that the straightforward binary Pauli algorithm from Section 1.4 is optimal in this setting.

Theorem 3.2.7. *Consider a tomography algorithm in the nonadaptive measurement model which outputs an estimate $\hat{\rho} \in \mathsf{D}(d)$ such that $\|\hat{\rho} - \rho\|_1 \leq \epsilon$ with at least a constant probability of success, for any unknown state $\rho \in \mathsf{D}(d)$. Suppose further that the each measurement is restricted to having a constant number of outcomes. Then it must hold that the algorithm uses $n = \Omega\left(d^4/\epsilon^2\right)$ samples of the state.*

We proceed in a similar fashion to the previous section. Let $\mathcal{M} = \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \cdots \otimes \mathcal{M}_n$ be the measurement used by the tomography algorithm, where each $\mathcal{M}_k$ is a constant-outcome measurement on operators in $\mathsf{D}(d)$. In other words, for every $k \in [n]$ there exists an $L = O(1)$ such that for any density matrix $\rho \in \mathsf{D}(d)$,

$$\mathcal{M}_k : \rho \mapsto \sum_{j=1}^{L} \operatorname{Tr}\left(M_j^{(k)} \rho\right) |j\rangle\langle j|$$

for some set of measurement operators $\{M_1^{(k)}, \ldots, M_L^{(k)}\}$. As explained in the previous section, it will suffice to bound $I(\boldsymbol{U} : \boldsymbol{z}_k)$ for each $k \in [n]$, where $\boldsymbol{z}_k$ is the outcome corresponding to the measurement $\mathcal{M}_k$ on the state $\rho_{\epsilon,\boldsymbol{U}}$ and $\boldsymbol{U}$ is a Haar random unitary in $\mathbb{U}(d)$. The inequality (3.10) from the more general nonadaptive case continues to hold:

$$I(\boldsymbol{U} : \boldsymbol{z}_k) \leq \log(e) \left[\mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{\boldsymbol{z}_k}} \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \frac{p_{\boldsymbol{z}_k|\boldsymbol{U}}(\boldsymbol{z})^2}{p_{\boldsymbol{z}_k}(\boldsymbol{z})^2} - 1 \right]$$

and the result follows upon bounding the right-hand side by $O(\epsilon^2/d^2)$ rather than $O(\epsilon^2/d)$ i.e., we would like to show that

$$\mathop{\mathbb{E}}_{\boldsymbol{z} \sim p_{\boldsymbol{z}_k}} \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \frac{p_{\boldsymbol{z}_k|U}(\boldsymbol{z})^2}{p_{\boldsymbol{z}_k}(\boldsymbol{z})^2} = \sum_z \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \frac{p_{\boldsymbol{z}_k|\boldsymbol{U}}(z)^2}{p_{\boldsymbol{z}_k}(z)} \tag{3.14}$$

is at most $1 + O(\epsilon^2/d^2)$ when we know that the sum on the right-hand side consists of $O(1)$ terms, for then the mutual information $I(\boldsymbol{U} : \boldsymbol{z}) = \sum_{k=1}^{n} I(\boldsymbol{U} : \boldsymbol{z}_k) = o(d^2)$ unless $n = \Omega(d^4/\epsilon^2)$. In summary, the result follows upon showing the following proposition.

Proposition 3.2.8. *Let $\boldsymbol{z}$ be a random variable corresponding to the outcome of a* constant-outcome *measurement $\mathcal{M}$ performed on the state $\rho_{\epsilon,\boldsymbol{U}}$, where $\boldsymbol{U}$ is Haar-random and $\rho_{\epsilon,U}$ is as in (3.1), for each $U \in \mathbb{U}(d)$. It holds that*

$$\ln(2) \times I(\boldsymbol{U} : \boldsymbol{z}) \leq \mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \chi^2(p_{\boldsymbol{z}|\boldsymbol{U}} \parallel p_{\boldsymbol{z}}) = O\left(\frac{\epsilon^2}{d^2}\right).$$

We know from the first identity in Lemma 3.2.6 that $p_{\boldsymbol{z}_k}(z) = w_k(z)$, where $w_k(z) := \mathrm{Tr}(M_z^{(k)}/d)$ and $M_z^{(k)}$ is the measurement operator corresponding to the outcome z for the k^{th} measurement. The desired upper bound on the right-hand side of (3.14) then follows from the next lemma, which is an alternative upper bound on the second expectation appearing in Lemma 3.2.6.

Lemma 3.2.9. *Let $d > 1$ be a positive integer, $M \in \mathsf{Psd}(d)$, $0 \preceq M \preceq \mathbb{1}$ be a measurement operator, and $\rho_{\epsilon,U}$ be defined as in (3.1), for each $U \in \mathbb{U}(d)$ and $\epsilon \in (0,1)$. Also, let $w = \mathrm{Tr}(M)/d$. It holds that*

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

Proof. We found in the proof of Lemma 3.2.6 (see the left-hand side of (3.13)) that

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

We may rewrite the right-hand side of the above as

$$w^2 + \frac{\epsilon^2 \mathrm{Tr}(M^2)}{d(d^2-1)} - \frac{\epsilon^2 w^2}{d^2-1} \leq w^2 + \frac{\epsilon^2 w}{d^2-1}$$

where the inequality follows since $\mathrm{Tr}(M^2) \leq \mathrm{Tr}(M)$ for $0 \preceq M \preceq \mathbb{1}$ and the third term in the left-hand side is at most zero. □

Note that $\sum_z p_{\boldsymbol{z}_k}(z) = \sum_z w_k(z) = 1$. The inequality in the lemma above along with the fact that the number of outcomes is $L = O(1)$ entails that the right-hand side of (3.14) is at most

$$\sum_{z=1}^{L} \frac{w_k(z)^2 + \epsilon^2 w_k(z)/(d^2-1)}{w_k(z)} = 1 + O(\epsilon^2/d^2)$$

as desired, completing the proof of Theorem 3.2.7.

Chapter 4

Lower bounds with adaptive measurements

We have previously seen that by taking into account restrictions on measurements, it is possible to derive stronger lower bounds on tomography than that obtained by a direct application of Holevo's theorem. Specifically, we were able to show matching lower bounds on tomography in the nonadaptive measurement setting for both constant-outcome measurements and arbitrary POVMs. In this chapter we will consider two different kinds of restriction on the measurements.

First, we examine the case where there is some fixed subset of the measurements which can be used to adapt future measurements. For this scenario, we are able to show a lower bound which matches the nonadaptive case, so long as the number of measurements which can be adapted on is not $\Omega(d^2/\epsilon^2)$.

We then allow measurements to adapt on arbitrarily many of the previous outcomes, so long as they are selected from a fixed set of up to $\exp(O(d))$ measurements. This is large enough to encompass all possible measurements which can be efficiently performed on a quantum computer with a fixed universal gate set.

Finally, we apply the above method to derive lower bounds on a different task called *classical shadows*, where the goal is to predict the expectation of a number of observables using few copies of the state. Namely, we show that the matching lower bound due to Huang, Kueng, and Preskill [22] is robust to adaptively selected measurements, so long as one restricts their attention to efficient quantum computation. Such an assumption is particularly relevant in this context, since classical shadows are thought to be a useful tool for proposed near-term applications of quantum computing [33]. It could be said that any advantage for adaptivity lying outside the scope we consider (i.e., requiring $\mathrm{e}^{\omega(d)}$ mea-

surement settings) would undermine the value of classical shadows in a practical context, since it could not be realized using a quantum device with a realistic number of operations, growing polynomially with the number of qubits.

4.1 Measurements with limited adaptivity

Suppose that a tomography algorithm uses only some past measurement outcomes to adapt future ones. For example, an experiment may begin with a preliminary phase in which measurements are performed fully adaptively, but then transition to a second phase in which measurements depend only on the outcomes obtained in the first one. (The proposal considered in Mahler et al. [27] meets this criterion, for example.) In this section we prove a lower bound on tomography with adaptive measurements when the algorithm adapts on just a fixed subset of the previously observed outcomes, which is not too large. More precisely, suppose there are n measurements made in total, which is equal to the number of samples/registers prepared in the state ρ in this setting. Then we assume that there is a subset of $S \subseteq [n]$ of size at most r such that, for each $i \in [n]$, the i^{th} measurement may depend only on the outcomes obtained from the registers corresponding to the measurements $[i-1] \cap S$. We find that adaptivity does not offer any advantage over independent measurements in terms of the worst-case sample complexity in such scenarios.

Theorem 4.1.1. *Consider a tomography algorithm in the adaptive measurement model which outputs an estimate $\hat{\rho} \in \mathsf{D}(d)$ such that $\|\hat{\rho} - \rho\|_1 \leq \epsilon$ with at least a constant probability of success, for any unknown state $\rho \in \mathsf{D}(d)$. Suppose further that each choice of measurement is independent of all but a fixed set of up to $r = o(d^2/\epsilon^2)$ outcomes. Then it must hold that the algorithm uses $n = \Omega(d^3/\epsilon^2)$ samples of the state.*

We state at the outset that we find this theorem follows from suitable adjustments to the analysis performed by Bubeck et al. [13] in the context of lower bounds for a different problem, known as quantum property testing. We will highlight where we borrow ideas for our proof. Compared to the independent case, the key difference when the measurements are allowed to be adaptive is that the subadditivity property of mutual information no longer holds. In other words, if $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are the measurement outcomes obtained upon measuring $\rho_{\boldsymbol{x}}^{\otimes n}$, it may not necessarily be the case that $I(\boldsymbol{x} : \boldsymbol{y}) \leq \sum_{k=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_k)$ as in the nonadaptive setting. We may however appeal to the chain rule for mutual information (see Fact 1.3.3), $I(\boldsymbol{x} : \boldsymbol{y}) = \sum_{k=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_k | \boldsymbol{y}_{k-1}, \ldots, \boldsymbol{y}_1)$ which always holds for the random variables $\boldsymbol{x}$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$.

If it were the case that each of the n conditional mutual information terms in the chain rule were $O(\epsilon^2/d)$, then we would be able to recover the $\Omega(d^3/\epsilon^2)$ lower bound from

Theorem 3.2.4 in the independent case. Unfortunately we do not know how to show this for the Haar-random ensemble we have been considering thus far. Instead, our proof technique – inspired by Ref. [13] – relies on concentration inequalities to attempt to control the deviation of the conditional mutual information terms away from the case where there is no adaptivity. (See Lemma 4.1.3.) This leads to excess terms which scale exponentially with r, the number of terms adapted on, unless $r \leq o(d^2/\epsilon^2)$. At present, this shortcoming seems to be an artifact of the strongest concentration inequality we are able to show, which takes the form of a subexponential tail. If the concentration results were any tighter, the approach would lead to an unconditional separation between tomography with and without entangled measurements, analogous to the separation shown in Ref. [13] for the task of quantum property testing.

We leave open the possibility of improvements in the relevant concentration inequalities we derive from the above lemma. For now, let us make a final remark on the relationship between the approach considered in this section and the lower bounds we will present in Section 4.2. There, the bounds apply to measurements which can be fully adaptive, so long as there is a fixed set of measurements which can be performed that is not too large. In particular, the tail bound (Lemma 4.1.3) shown in this section will be applied to rule out any advantage in adaptivity for a fixed set of up to $\exp(O(d))$ measurements. Let us now define the function for which we will show the aforementioned tail bound.

Definition 4.1.2. Let $\Xi(d)$ be the set of all finite-outcome POVMs acting on operators in $\mathsf{D}(d)$. We define the function $\mathrm{X} : \Xi(d) \times \mathbb{U}(d) \to \mathbb{R}$ by

$$\mathrm{X}(\mathcal{M}, U) := \chi^2(p_{\boldsymbol{z}|U} \parallel w)$$

for any $U \in \mathbb{U}(d)$ and $\mathcal{M} \in \Xi(d)$, where, $p_{\boldsymbol{z}|U}$ is the measurement outcome distribution from applying $\mathcal{M}$ to the state $\rho_{\epsilon,U}$, and $w = \mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}}\, p_{\boldsymbol{z}|\boldsymbol{U}}$.

We have $\|\mathrm{X}\|_\infty \leq \epsilon^2$ since for any $\mathcal{M}$ with measurement operators $\{M_z\}_z$ and $U \in \mathbb{U}(d)$ one has by definition

$$|\mathrm{X}(\mathcal{M}, U)| = \mathop{\mathbb{E}}_{\boldsymbol{z}'\sim p_{\boldsymbol{z}}} \left(\frac{\mathrm{Tr}(M_{\boldsymbol{z}'}\rho_{\epsilon,U}) - w(\boldsymbol{z}')}{w(\boldsymbol{z}')} \right)^2 = \epsilon^2 \mathop{\mathbb{E}}_{\boldsymbol{z}'\sim p_{\boldsymbol{z}}} \left(\frac{\mathrm{Tr}(M_{\boldsymbol{z}'}\sigma) - w(\boldsymbol{z}')}{w(\boldsymbol{z}')} \right)^2 \leq \epsilon^2$$

where $w(z) = \mathrm{Tr}(M_z)/d$ and we used the definition of $\rho_{\epsilon,U}$ to write the expression in terms of the quantum state σ. (Note that $0 \leq w(z), \mathrm{Tr}(M_z\sigma) \leq 2w(z)$.)

We next turn to the tail inequality which will be used in this section and in Section 4.2 to derive lower bounds in adaptive measurement scenarios. We will only require the case of arbitrary (finite-outcome) POVMs for the proof of Theorem 4.1.1, but we will also use the constant-outcome case in later sections.

Lemma 4.1.3 (Chi-squared tail bound)**.** *For any finite-outcome POVM $\mathcal{M}$ it holds that*

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{X}(\mathcal{M},\boldsymbol{U}) > \gamma + t\right] \leq 2\exp\left(-\frac{Cd^2t}{\epsilon^2}\right) \tag{4.1}$$

where $\gamma := c\epsilon^2/d$ and c, C are universal constants. Furthermore, if $\mathcal{M}$ is restricted to having a constant number of outcomes then the inequality holds with $\gamma := a\epsilon^2/d^2$ for some universal constant a.

Proof. The lemma is analogous to Lemma 7.6 in Ref. [13], for example, and so we proceed along similar lines considering first the case where $\mathcal{M}$ has more than a constant number of outcomes. We are trying to prove a subexponential tail on the random variable $\mathrm{X}(\mathcal{M},\boldsymbol{U}) - c\epsilon^2/d$, where $\boldsymbol{U}$ is Haar-random. It will suffice to show that the function which acts on any $U \in \mathbb{U}(d)$ as

$$f : U \mapsto \sqrt{\mathrm{X}(\mathcal{M},U)} - \mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} \sqrt{\mathrm{X}(\mathcal{M},\boldsymbol{V})}$$

has a tail like $2\exp(-\Omega(d^2t^2/\epsilon^2))$, for U selected Haar-randomly. Let us explain the reason for this. Note that

$$\mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} \sqrt{\mathrm{X}(\mathcal{M},\boldsymbol{V})} \leq \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{V}\sim\mathrm{Haar}} \mathrm{X}(\mathcal{M},\boldsymbol{V})} = \frac{c'\epsilon}{\sqrt{d}}$$

where the inequality uses the concavity of the square root and the second relation follows from Proposition 3.2.5. Furthermore, the inequality $c\epsilon^2/d + t \geq \left(\sqrt{c\epsilon^2/d} + \sqrt{t}\right)^2/2$ for any $t > 0$ entails that

$$\Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\mathrm{X}(\mathcal{M},\boldsymbol{U}) > \frac{c\epsilon^2}{d} + t\right] \leq \Pr_{\boldsymbol{U}\sim\mathrm{Haar}}\left[\sqrt{\mathrm{X}(\mathcal{M},\boldsymbol{U})} > \epsilon\sqrt{\frac{c}{2d}} + \sqrt{\frac{t}{2}}\right]$$

so that, by choosing $c = 2(c')^2$ we find f having a tail of $2\exp(-\Omega(d^2t^2/\epsilon^2))$ indeed gives us

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

for some universal constant C, as required.

To arrive at the desired concentration inequality for f we can invoke Lemma 3.1.2, according to which it suffices to show that f is $O(\epsilon/\sqrt{d})$-Lipschitz. Let $\boldsymbol{z}$ be the outcome

from measuring $\rho_{\epsilon,\boldsymbol{U}}$ with $\mathcal{M}$ and having conditional distribution $p_{\boldsymbol{z}|U}$ given $\boldsymbol{U} = U$. Also define the distribution $w = \mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} p_{\boldsymbol{z}|\boldsymbol{U}}$. For arbitrary $U, V \in \mathbb{U}(d)$, consider the difference

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

In the above, the second line follows from the definition of $\mathrm{X}(\mathcal{M}, U)$ as well as the chi-squared divergence between two discrete distributions, and the third line follows from triangle inequality applied to the L^2 norm. Let $\{M_z\}_z$ be the set of measurement operators corresponding to $\mathcal{M}$. We have that $p_{\boldsymbol{z}|U}(z) = \mathrm{Tr}\left(M_z \rho_{\epsilon,U}\right)$ and $w(z) = \mathrm{Tr}(M_z)/d$ from Lemma 3.2.6. Recalling the definition of $\rho_{\epsilon,U}$ from (3.1) we may substitute into the right-hand side of the above inequality to arrive at the upper bound

$$|f(U) - f(V)| \leq \frac{2\epsilon}{d}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{z}\sim w} \frac{1}{w(\boldsymbol{z})^2}\mathrm{Tr}\left(M_{\boldsymbol{z}}(UQU^\dagger - VQV^\dagger)\right)^2}.$$

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

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{z}\sim w} \frac{1}{w(\boldsymbol{z})^2}\mathrm{Tr}\left(M_{\boldsymbol{z}} W D W^\dagger\right)^2 &= d^2 \mathop{\mathbb{E}}_{\boldsymbol{z}\sim w} \mathrm{Tr}\left(\left(\frac{W^\dagger M_{\boldsymbol{z}} W}{w(\boldsymbol{z})d}\right) D\right)^2 \\
&\leq d^2 \mathop{\mathbb{E}}_{\boldsymbol{z}\sim w} \mathrm{Tr}\left(\left(\frac{W^\dagger M_{\boldsymbol{z}} W}{w(\boldsymbol{z})d}\right) D^2\right) \\
&= d \sum_z \mathrm{Tr}\left(W^\dagger M_z W D^2\right) \\
&= d\left\|UQU^\dagger - VQV^\dagger\right\|_{\mathrm{F}}^2
\end{aligned}$$

where in the second line we used the fact 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 final line we used the fact that the measurement operators for the different outcomes z sum to identity.

Finally, we can use the matrix inequality $\|AB\|_{\mathrm{F}} \leq \|A\| \|B\|_{\mathrm{F}}$ to deduce that

$$\begin{aligned}\left\|UQU^\dagger - VQV^\dagger\right\|_{\mathrm{F}} &= \frac{1}{2}\left\|(U+V)Q(U-V)^\dagger + (U-V)Q(U+V)^\dagger\right\|_{\mathrm{F}} \\ &\leq \left\|(U+V)Q(U-V)^\dagger\right\|_{\mathrm{F}} \\ &\leq (\|UQ\| + \|VQ\|)\|U-V\|_{\mathrm{F}} \\ &\leq 2\|U-V\|_{\mathrm{F}}. \end{aligned} \tag{4.2}$$

The proof in the constant-outcome case is identical, except using the fact that the expectation is then $\mathbb{E}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathrm{X}(\mathcal{M}, \boldsymbol{U}) = O(\epsilon^2/d^2)$ in accordance with Proposition 3.2.8. □

Let us return to proving Theorem 4.1.1. We begin by introducing some short-hand notation which will help when analyzing the conditioning of measurements on previous outcomes. Let $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ be the measurement outcomes from performing a sequence of adaptive measurements on n identical copies of the state $\rho_{\epsilon,\boldsymbol{U}}$, where $\boldsymbol{U}$ is a Haar-random unitary operator and $\rho_{\epsilon,U}$ is defined as in previous sections to be of the form (3.1) for any $U \in \mathbb{U}(d)$. We say that the random variable $\boldsymbol{z}_i$ can be *adapted on* if there is some choice of measurement which depends on the outcome $\boldsymbol{z}_i$. Let $\boldsymbol{z}_{<k}$ denote the random variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{k-1}$ and $z_{<k}$ a possible sequence of values $z_1, \ldots, z_{k-1}$.

By chain rule for mutual information we have that

$$I(\boldsymbol{U} : \boldsymbol{z}) = \sum_{k=1}^{n} I(\boldsymbol{U} : \boldsymbol{z}_k | \boldsymbol{z}_{<k}). \tag{4.3}$$

From now on we will focus on a specific term k in the sum in (4.3), since it suffices to show that each term is $O(\epsilon^2/d)$. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{r-1}$ be the subset of outcomes which can be adapted on out of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{k-1}$ and write $\boldsymbol{y}_r := \boldsymbol{z}_k$. It follows from Fact 1.3.2 that $I(\boldsymbol{U} : \boldsymbol{z}_k | \boldsymbol{z}_{<k}) \leq I(\boldsymbol{U} : \boldsymbol{z}_k | \boldsymbol{y}_{<r}) =: I(\boldsymbol{U} : \boldsymbol{y}_r | \boldsymbol{y}_{<r})$.

For each $j \in [r]$ and $y_{<j}$ we have that the conditional distribution for the outcome $\boldsymbol{y}_j$ satisfies

$$p_{\boldsymbol{y}_j | y_{<j}, U}(y_j) = \mathrm{Tr}\left(M_{y_j}^{y_{<j}} \rho_{\epsilon,U}\right)$$

for some set of measurement operators $\{M_{y_j}^{y_{<j}}\}_{y_j}$. (See Section 1.4 for a description of the adaptive measurement model.) We will denote this POVM by $\mathcal{M}^{y_{<j}}$. Also, define

$$w^{y_{<j}} := \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} p_{\boldsymbol{y}_j | y_{<j}, \boldsymbol{U}} = \mathrm{Tr}(M_{y_j}^{y_{<j}})/d,$$

where the final equality follows from the first expectation in Lemma 3.2.6. Finally, define the distribution $w^{<r}$ over outcomes $y_{<r}$ with probabilities $w^{<r}(y_{<r}) = \prod_{j=1}^{r-1} w^{y_{<j}}(y_j)$.

We can now proceed with bounding the conditional mutual information term $I(\boldsymbol{U} : \boldsymbol{y}_r|\boldsymbol{y}_{<r})$. Recalling the upper bound on the mutual information from Lemma 3.2.1 we have

$$\begin{aligned}
I(\boldsymbol{U} : \boldsymbol{y}_r|\boldsymbol{y}_{<r}) &\leq \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}} \mathop{\mathbb{E}}_{\boldsymbol{U}|\boldsymbol{y}_{<r}} \chi^2(p_{\boldsymbol{y}_r|\boldsymbol{y}_{<r},\boldsymbol{U}} \parallel w^{\boldsymbol{y}_{<r}}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}} \mathop{\mathbb{E}}_{\boldsymbol{U}|\boldsymbol{y}_{<r}} \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}|\boldsymbol{U}} \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \sum_{y_{<r}} \prod_{j=1}^{r-1} p_{\boldsymbol{y}_j|y_{<j},\boldsymbol{U}}(y_j)\ \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U}) \\
&= \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}\sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j},\boldsymbol{U}}(\boldsymbol{y}_j)}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)}\ \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U}) \qquad (4.4)
\end{aligned}$$

where in the fourth line we use the fact that for fixed U we have that the conditional distribution over prior measurement outcomes $y_{<r}$ satisfies

$$p_{\boldsymbol{y}_{<r}|U}(y_{<r}) = \prod_{j=1}^{r-1} p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j},U}(y_j),$$

and in the final line we factor out the probability $w^{<r}(y_{<r})$ from each term in the sum. Also note that we are redefining the random variable $\boldsymbol{y}_{<r}$ in the final line by writing the expectation over $\boldsymbol{y}_{<r} \sim w^{<r}$, for ease of notation. We will now try to bound this expectation using the tail bound in Lemma 4.1.3.

Let $\mathbf{1}_E$ be the indicator function for the event E where the random variable $\mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U})$ is at most $c\epsilon^2/d + \tau$ for c the universal constant appearing in Lemma 4.1.3 and $\tau > 0$ to be specified later. Let $\overline{E}$ be the complement of this event. The expectation in (4.4) can be written

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}\sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j},\boldsymbol{U}}(\boldsymbol{y}_j)}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)}\ \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<r}}, \boldsymbol{U})\,(\mathbf{1}_E + \mathbf{1}_{\overline{E}}) \\
&\leq \tau + \frac{c\epsilon^2}{d} + \epsilon^2 \mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}\sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{Y_j|Y_{<j},U}(\boldsymbol{y}_j|\boldsymbol{y}_{<j}, \boldsymbol{U})}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)} \mathbf{1}_{\overline{E}}
\end{aligned}$$

$$\leq \tau + \frac{c\epsilon^2}{d} + 2\epsilon^2 \exp\left(-\Theta\left(\frac{d^2\tau}{\epsilon^2}\right)\right) \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\text{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}\sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j},\boldsymbol{U}}(\boldsymbol{y}_j)^2}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)^2}}. \tag{4.5}$$

Here, the second line uses the fact that $\sum_{y_{<r}} p_{\boldsymbol{y}_{<r}|U}(y_{<r}) = 1$ for any U along with the bound $\|\mathrm{X}\|_\infty \leq \epsilon^2$, and the last line follows from Cauchy-Schwarz inequality and Lemma 4.1.3, which is our tail bound on $\mathrm{X}(\mathcal{M}, \cdot)$ for any fixed measurement operator $\mathcal{M}$. It remains to show a suitably small upper bound on the expectation appearing in the square root in (4.5). To accomplish this, let us first define an ancillary random variable $\boldsymbol{K}$ whose moments we can bound more easily, and then relate the quantity in our expectation to this newly defined random variable.

Lemma 4.1.4. *Fix a constant $\alpha > 0$ and define the random variable $\boldsymbol{K}(\mathcal{M}, \alpha) = 1 + \alpha \mathrm{X}(\mathcal{M}, \boldsymbol{U})$ for any POVM $\mathcal{M}$. For $p \leq o(d^2/\epsilon^2)$ the p^{th} moment of $\boldsymbol{K}(\mathcal{M}, \alpha)$ satisfies*

$$\mathbb{E}\boldsymbol{K}(\mathcal{M}, \alpha)^p \leq C \exp\left(\frac{C' p\epsilon^2}{d}\right) + o(1) \tag{4.6}$$

for some universal constants C, C', where the expectation is taken with respect to the Haar-random distribution over the random variable $\boldsymbol{U}$.

Proof. We will work with the cumulative distribution function $F(x) := \Pr[\mathrm{X}(\mathcal{M}, \boldsymbol{U}) \leq x]$ as well as the probability density $p(x) := F'(x)$. Let $G(x) := (1 + \alpha x)^p$ be a function over the reals, which is nondecreasing for all $x \geq 0$ and $p \geq 1$. Also define $x_0 := 2c\epsilon^2/d$ where c is the constant appearing in Lemma 4.1.3 and $x_1 := \epsilon^2$. We can then write our expectation as

$$\begin{aligned}
\mathbb{E}\,\boldsymbol{K}(\mathcal{M}, \alpha)^p &= \mathbb{E}\, G(\mathrm{X}(\mathcal{M}, \boldsymbol{U})) \\
&= \int_0^{x_1} G(x) p(x) \mathrm{d}x \\
&\leq G(x_0) + \int_{x_0}^{x_1} G(x) p(x) \mathrm{d}x
\end{aligned}$$

where the limits of the integral in the first line come from the fact that $\Pr[0 \leq \mathrm{X}(\mathcal{M}, \boldsymbol{U}) \leq x_1] = 1$ and the inequality follows by observing that G is nondecreasing on the interval $[0, \infty)$. Let us now perform integration by parts on the integral and use the fact that $F(x_1) = 1$ to deduce that the right-hand side of the above is at most

$$G(x_0) + G(x_1) - G(x_0)F(x_0) - \int_{x_0}^{x_1} G'(x) F(x) \mathrm{d}x$$

$$
\begin{aligned}
&\leq G(x_0) + G(x_1) + \int_{x_0}^{x_1} G'(x)(1 - F(x))\mathrm{d}x - G(x_1) + G(x_0) \\
&= 2G(x_0) + \int_{x_0}^{x_1} G'(x) \Pr[\mathrm{X}(\mathcal{M}, \boldsymbol{U}) > x]\mathrm{d}x \\
&\leq 2G(x_0) + 2\int_{x_0}^{x_1} G'(x) \exp\left(-\frac{Cd^2(x - c\epsilon^2/d)}{\epsilon^2}\right)\mathrm{d}x \qquad (4.7) \\
&\leq 2G(x_0) + 2\int_{x_0}^{x_1} G'(x) \exp\left(-\frac{Cd^2 x}{2\epsilon^2}\right)\mathrm{d}x \qquad (4.8) \\
&\leq 2G(x_0) + 2\int_{0}^{\infty} p\alpha\,(1 + \alpha x)^{p-1} \exp\left(-\frac{Cd^2 x}{2\epsilon^2}\right)\mathrm{d}x \qquad (4.9) \\
&=: 2G(x_0) + 2p\int_{0}^{\infty} (1 + s)^{p-1} \exp\left(-\frac{c'd^2 s}{\epsilon^2}\right)\mathrm{d}s \qquad (4.10)
\end{aligned}
$$

where in the second line we neglect a negative term, in (4.7) we use Lemma 4.1.3, in (4.8) we observe $x - c\epsilon^2/d \geq x/2$ for all $x \geq x_0 = 2c\epsilon^2/d$, in (4.9) we substitute for $G'(x)$ and increase the limits of the integral (as the integrand is nonnegative) and in the final line we make the substitution $s := \alpha x$. Letting $a = c'd^2/\epsilon^2$, we appeal to the integral identity

$$
\int_0^\infty (1 + s)^{p-1} e^{-as}\mathrm{d}s = a^{-p} e^a \Gamma(p, a)
$$

where $\Gamma(\cdot, \cdot)$ is the incomplete Gamma function. (See Appendix B.1). Using the fact that $\Gamma(p, a) = a^{p-1}e^{-a}(1 + o(1))$ for $p = o(a)$ (and $a = \Theta(d^2/\epsilon^2)$ so this is true by assumption), we have that the expression in (4.10) is at most

$$
\begin{aligned}
2G(x_0) + 2pa^{-1}(1 + o(1)) &= 2\left(1 + \frac{2c\alpha\epsilon^2}{d}\right)^p + \frac{2p\epsilon^2}{d^2}(1 + o(1)) \\
&\leq C \exp\left(\frac{C'p\epsilon^2}{d}\right) + o(1)
\end{aligned}
$$

so long as $p = o(d^2/\epsilon^2)$ and defining C and C' appropriately, as required. □

Now we present the lemma which relates our expectation value of interest from (4.5) to the random variable defined in the lemma above.

Lemma 4.1.5. *The following upper bound on the expectation in (4.5) holds for some POVM $\mathcal{M}$ and absolute constant $\alpha > 0$:*

$$
\mathop{\mathbb{E}}_{\boldsymbol{U}\sim\mathrm{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}\sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j},\boldsymbol{U}}(\boldsymbol{y}_j)^2}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)^2} \leq \mathbb{E}\,\boldsymbol{K}(\mathcal{M}, \alpha)^{r-1} \qquad (4.11)
$$

where $\boldsymbol{K}(\mathcal{M}, \alpha)$ is defined as in Lemma 4.1.4.

Proof. Similar steps for this proof are also used in Ref. [13] in the context of a lemma necessary for their lower bound on quantum identity testing. First define

$$g_U^{y_{<j}}(y_j) = \frac{p_{\boldsymbol{y}_j|y_{<j},U}(y_j)}{w^{y_{<j}}(y_j)} - 1$$

For any $U \in \mathbb{U}(d)$, $j \in \{1, \ldots, r-1\}$, fixed $y_{<j}$, and $\ell \geq 2$ we can expand using generalized binomial coefficients to find

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{y}_j \sim w^{y_{<j}}} \left(1 + g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j)\right)^{\ell} &= 1 + \mathop{\mathbb{E}}_{\boldsymbol{y}_j \sim w^{y_{<j}}} \sum_{p=2}^{\infty} \binom{\ell}{p} \left(g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j)\right)^p \\
&\leq 1 + 2^{\ell} \mathop{\mathbb{E}}_{\boldsymbol{y}_j \sim w^{y_{<j}}} g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j)^2 \\
&= 1 + 2^{\ell} \mathrm{X}(\mathcal{M}^{y_{<j}}, \boldsymbol{U}) \\
&=: \boldsymbol{K}(\mathcal{M}^{y_{<j}}, 2^{\ell})
\end{aligned}$$

where in the first line we used the fact that $\mathbb{E}_{\boldsymbol{y}_j \sim w^{y_{<j}}} g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j) = 0$ and in the last line we are using the fact that $g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j)^p \leq g_{\boldsymbol{U}}^{y_{<j}}(\boldsymbol{y}_j)^2$ for all $p \geq 2$. Hence, the left-hand side of (4.11) is upper bounded by

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r-1}} \prod_{j=1}^{r-2} \left(1 + g_{\boldsymbol{U}}^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)\right)^2 \boldsymbol{K}(\mathcal{M}^{\boldsymbol{y}_{<r-1}}, 4) \leq &\left[\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r-1}} \prod_{j=1}^{r-2} \left(1 + g_{\boldsymbol{U}}^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)\right)^{\frac{2(r-1)}{r-2}}\right]^{\frac{r-2}{r-1}} \\
&\times \left[\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r-1}} \boldsymbol{K}(\mathcal{M}^{\boldsymbol{y}_{<r-1}}, 4)^{r-1}\right]^{\frac{1}{r-1}}
\end{aligned}$$

where we have applied Hölder's inequality $\mathbb{E}\,|\boldsymbol{X}\boldsymbol{Z}| \leq (\mathbb{E}\,|\boldsymbol{X}|^p)^{1/p} (\mathbb{E}\,|\boldsymbol{Z}|^q)^{1/q}$ with $p = (r-1)/(r-2)$ and $q = r-1$ and $\boldsymbol{y}_{<r-1}$ has distribution $w^{<r-1}$ with probabilities $w^{<r-1}(y_{<r-1}) = \prod_{j=1}^{r-2} w^{y_{<j}}(y_j)$. We will now make an assumption that will be justified at the end of this proof: assume that the term being raised to the power of $(r-2)/(r-1)$ is at least 1. Then, we may raise that term to the power of $(r-1)/(r-2)$ only to increase it, which enables us to proceed by a recursive argument. Let us introduce some further notation that will clarify how to apply the recursion. Define

$$\boldsymbol{S}_a := \prod_{j=1}^{a-1} (1 + g_{\boldsymbol{U}}^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)).$$

Furthermore, let $\ell_i = 2\left(\frac{r-1}{r-2}\right)^i$ for every $i \in \mathbb{Z}_+$. Substituting, we have the upper bound

$$\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}} \boldsymbol{S}_r^{\ell_0} \le \left[\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r-1}} \boldsymbol{K}(\mathcal{M}^{\boldsymbol{y}_{<r-1}}, 2^{\ell_0})^{r-1}\right]^{\frac{1}{r-1}} \mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r-1}} \boldsymbol{S}_{r-1}^{\ell_1}.$$

Applying this argument recursively to the moments of the $\boldsymbol{S}_a$ random variables (and assuming at each step that they are at least 1) we arrive at

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r}} \boldsymbol{S}_r^{\ell_0} &\le \prod_{i=1}^{r-1} \left[\mathop{\mathbb{E}}_{\boldsymbol{U}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<i}} \boldsymbol{K}(\mathcal{M}^{\boldsymbol{y}_{<i}}, 2^{\ell_{r-1-i}})^{r-1-i}\right]^{\frac{1}{r-1}} \\ &\le \sup_{i, y_{<i}} \boldsymbol{K}(\mathcal{M}^{y_{<i}}, 2^{\ell_{r-1-i}})^{r-1}\end{aligned} \tag{4.12}$$

where in going from the first to the second line we first took the supremum over all terms i in the product as an upper bound and then used the fact that the expectation for the i^{th} term is over a product distribution of the random variables $\boldsymbol{U}, \boldsymbol{y}_{<i}$ to take the supremum over all terms $y_{<i}$ as an upper bound. It remains to show that the argument ℓ_{r-1-i} is bounded by some absolute constant for all $i \in \{1, \ldots, r-1\}$. To see this, note that $(r-1)/(r-2) = \left(1 + \frac{1}{r-2}\right)$ and $r - 1 - i \le r - 2$ for every $i \in \{1, \ldots, r-1\}$. Hence, $\ell_{r-1-i} \le 2\left(1 + \frac{1}{r-2}\right)^{r-2} \le 2\mathrm{e}$ so that the lemma holds with $\alpha = 4^{\mathrm{e}}$.

We return now to the assumption we are making at each step of the recursion that the moments of $\boldsymbol{S}_a$ are lower bounded by 1. Suppose that this assumption fails at some step $1 \le t \le r-1$ of the recursion. We then have an upper bound analogous to the right-hand side of (4.12), but with t terms in the product. Since $\boldsymbol{K}(\mathcal{M}^{y_{<i}}, \alpha)$ is certainly at least 1, raising each term in the product to the power of $(r-1)/t$ is an upper bound. We can then take the supremum over these t terms to arrive at the same result. □

We are now ready to prove Theorem 4.1.1.

Proof of Theorem 4.1.1. Lemmas 4.1.4 and 4.1.5 together imply the upper bound

$$\mathop{\mathbb{E}}_{\boldsymbol{U} \sim \text{Haar}} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<r} \sim w^{<r}} \prod_{j=1}^{r-1} \frac{p_{\boldsymbol{y}_j|\boldsymbol{y}_{<j}, \boldsymbol{U}}(\boldsymbol{y}_j)^2}{w^{\boldsymbol{y}_{<j}}(\boldsymbol{y}_j)^2} \le C \exp\left(\Theta(r\epsilon^2/d)\right) + o(1) \tag{4.13}$$

for some absolute constant C. Substituting into the right-hand side of (4.5) (which is our upper bound on the k^{th} conditional mutual information term) we get

$$I(\boldsymbol{U} : \boldsymbol{z}_k | \boldsymbol{z}_{<k}) \le \tau + O\left(\frac{\epsilon^2}{d}\right) + 2C\epsilon^2 \exp\left(-\Theta\left(\frac{d^2\tau}{\epsilon^2}\right) + \Theta\left(\frac{r\epsilon^2}{d}\right)\right)$$

for some constant C. We must show that for appropriately chosen τ, the right-hand side is not too large. Choosing

$$\tau = \frac{c\epsilon^2}{d} + \frac{c'\epsilon^2 \log(d)}{d^2}$$

for appropriately chosen constants c, c' and noting that $r = o(d^2/\epsilon^2)$ we get an upper bound of

$$I(\boldsymbol{U} : \boldsymbol{z}_k | \boldsymbol{z}_{<k}) \leq O\left(\frac{\epsilon^2}{d}\right).$$

Applying this upper bound to each term in the chain rule we obtain

$$I(\boldsymbol{U} : \boldsymbol{z}) \leq O\left(\frac{n\epsilon^2}{d}\right)$$

so that the proof of Theorem 4.1.1 is complete upon invoking Fano's inequality in the same fashion as the previous sections. □

4.2 Lower bounds for adaptive tomography with efficient measurements

Fix a set of m POVMs $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$. In this section, we claim the existence of a large ϵ-packing of mixed states which possesses the additional property of leading to highly uninformative measurement outcomes from each of the m measurements. The $\Omega(d^4/\log(d))$ lower bound for adaptive Pauli measurements from Ref. [16] is obtained using this mindset, except there the packing is constructed with the property that all possible binary Pauli measurements lead to outcome distributions that are close to uniform, for every state in the packing. In contrast, we do not enforce that the individual probabilities are close to some fixed value, but rather directly bound the magnitude of the χ^2 divergence (what we have been calling the function X). This leads to tight lower bounds agnostic to the measurements comprising our set.

Theorem 4.2.1. *Consider a tomography algorithm in the adaptive measurement model which outputs an estimate $\hat{\rho} \in \mathsf{D}(d)$ such that $\|\hat{\rho} - \rho\|_1 \leq \epsilon$ with at least a constant probability of success, for any unknown state $\rho \in \mathsf{D}(d)$. Suppose further that each measurement is selected from a fixed set of up to $\exp(O(d))$ different measurements. Then it must hold that the algorithm uses $n = \Omega(d^3/\epsilon^2)$ samples of the state.*

We also have the following extension of the above theorem.

Theorem 4.2.2. *Suppose the assumptions in Theorem 4.2.1 are satisfied, and further that each measurement has a constant number of outcomes. Then* $n = \Omega(d^4/(\epsilon^2 \log(d)))$ *samples are required.*

Combined with the following property of quantum circuits, Theorem 4.2.1 limits the potential for an improvement in the worst-case sample complexity using efficient, adaptive measurements. Here, we adopt the definition of quantum circuits which allows us to model POVMs besides unitary evolution followed by basis measurements[1].

Proposition 4.2.3. *Consider a q-qubit register in some unknown mixed state* $\rho \in \mathsf{D}(d)$*, where* $d = 2^q$ *and let* $\mathcal{G}$ *be a finite universal gate set comprised of a constant number of gates. The number of distinct measurements which can be implemented using a quantum circuit of* $\mathrm{poly}(q)$ *gates from* $\mathcal{G}$ *is* $O(\mathrm{poly}(q)^{\mathrm{poly}(q)}) = \exp(O(d))$*.*

One implication of the result in Theorem 4.2.1 is that the nonadaptive 2-design tomography protocol from Section 1.5 is optimal in the setting of efficient quantum computation. This can be seen from the fact that there exist unitary 2-designs that admit efficient implementations on a quantum circuit. For example, each element of the group of q-qubit Clifford circuits (forming a 3-design) can be implemented with $\mathrm{poly}(q)$ gates (see Corollary 9 in Ref. [2] for example). Another example of efficiently implementable measurements is adaptive d-outcome Pauli basis measurements on q qubits as considered in the recent work of Yu [41] where a $O(10^q/\epsilon^2)$ upper bound on tomography is shown. The result from this section yields a $\Omega(8^q/\epsilon^2)$ lower bound in that setting, improving the gap one obtains from Holevo's theorem which is $\Omega(4^q/\epsilon^2)$.

Proof of Theorem 4.2.1

Let us move on to the proof of the theorem. Since our intermediate goal is to construct an ϵ-packing of states using a probabilistic existence argument we will require a result analogous to Lemma 3.1.4, which we used to derive the previous packing of states.

Proposition 4.2.4. *There exists a universal constant c such that the following holds. Pick* $\epsilon \in (0, 1)$*, let* $d > 0$ *be a positive integer and let* $0 \leq N < \mathrm{e}^{cd^2}/4$ *be an integer. Consider a*

[1] A gate is a quantum *operation* (a CPTP map taking mixed states to other mixed states), and the universal gateset $\mathcal{G}$ can be used to enact any quantum operation from q to m qubits to within ε accuracy in *diamond norm*, with only polynomial overhead [38]. An example is the Clifford+T gateset, along with the erasure and ancilla gates to model the possibility of adding and tracing out qubits.

set of states $\{\rho_1, \rho_2, \ldots, \rho_N\} \subset \mathsf{D}(d)$ *where*

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

for each $i \in [N]$, $U_1, U_2, \ldots, U_N \in \mathbb{U}(d)$ *are arbitrary unitary operators, and* σ *is as in* (3.1). *For Haar-random* $\boldsymbol{U}$ *taking values in* $\mathbb{U}(d)$, *the probability that* $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$ *for any* $i \in [N]$ *is at most* $1/2$.

Proof. The proof is identical to that for Lemma 3.1.4. □

Next, let $\mathcal{K} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$ denote our set of m available measurements, and let us first consider the case where these can be arbitrary POVMs. We will invoke the tail inequality from Lemma 4.1.3 to claim that for a large fraction of unitaries in $\mathbb{U}(d)$, the measurement statistics from these POVMs are uninformative. Specifically, letting $\boldsymbol{y}_k$ denote the outcome from performing the measurement $\mathcal{M}_k$ on $\rho_{\epsilon,\boldsymbol{U}}$ for Haar-random $\boldsymbol{U}$, the lemma says that for some fixed distribution w, there are many unitary operators $U \in \mathbb{U}(d)$ such that $\chi^2(p_{\boldsymbol{y}_k|U} \parallel w)$ is small for every $k \in [m]$.

Lemma 4.2.5. *Fix a set of* m *POVMs* $\mathcal{K} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m\}$ *acting on operators in* $\mathsf{D}(d)$, *and define* $\mathrm{X}(\mathcal{M}, U)$ *as in Definition 4.1.2. Let* $\beta = \epsilon^2 \ln(6m)/(Cd^2)$, *where* C *is the absolute constant appearing in the tail in Lemma 4.1.3. For Haar-random* $\boldsymbol{U}$ *the probability that* $\mathrm{X}(\mathcal{M}, \boldsymbol{U}) \leq O(\epsilon^2/d) + \beta$ *for every measurement* $\mathcal{M} \in \mathcal{K}$ *is at least* $2/3$.

Proof. Applying union bound over the various measurements we find that the probability that there is some measurement $\mathcal{M} \in \mathcal{K}$ such that $\mathrm{X}(\mathcal{M}, \boldsymbol{U}) > c\epsilon^2/d + \beta$ is at most

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

where the inequality follows from the tail bound in Lemma 4.1.3 and c is as in that lemma. □

Combining Proposition 4.2.4 with Lemma 4.2.5 we can use a probabilistic existence argument to claim that there is a packing which is especially difficult to discriminate using any of the measurements in $\mathcal{K}$. This is the content of the proceeding corollary.

Corollary 4.2.6. *Define* $\mathrm{X}(\mathcal{M}, U)$, $\mathcal{K}$, *and* β *as in Lemma 4.2.5. There exists a set of* $N = \exp(\Omega(d^2))$ *mixed states* $\mathcal{S} = \{\rho_1, \ldots, \rho_N\} \subset \mathsf{D}(d)$ *of the form*

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

for some unitary operators $U_1, \ldots, U_N \in \mathbb{U}(d)$ *and* σ *as in* (3.1) *such that*

1. $\|\rho_i - \rho_j\|_1 > \epsilon/2$ *for every* $i, j \in [N], i \neq j$ *and*

2. $\mathrm{X}(\mathcal{M}, U_i) \leq O(\epsilon^2/d) + \beta$ *for every* $i \in [N]$ *and* $\mathcal{M} \in \mathcal{K}$.

Proof. The proof is similar to that of Corollary 3.1.6, except with an extra step. Suppose we have a set of $j \leq e^{cd^2 - \ln(4)}$ quantum states $\mathcal{S}_j = \{\rho_1, \ldots, \rho_j\}$ which are of the same form as in the statement of the corollary we are trying to prove, satisfying the two conditions with corresponding unitary operators $\mathcal{U}_j = \{U_1, \ldots, U_j\}$. From Proposition 4.2.4 as well as Lemma 4.2.5 we know that the probability of selecting a unitary $\boldsymbol{U}$ Haar-randomly such that $\mathcal{S}_{j+1} := \mathcal{S}_j \cup \rho_{\epsilon,\boldsymbol{U}}$ no longer satisfies either the first or the second condition is strictly less than one, where we have applied union bound on this event. To be precise, the probability that $\|\rho_{\epsilon,\boldsymbol{U}} - \rho_i\|_1 \leq \epsilon/2$ for some $i \in [j]$ *or* that $\mathrm{X}(\mathcal{M}, \boldsymbol{U}) > O(\epsilon^2/d) + \beta$ for some $\mathcal{M} \in \mathcal{K}$ is strictly less than one. Therefore, there must exist at least one such state, and the result follows by induction on j. □

We can now prove the lower bound in Theorem 4.2.1. Let $\mathcal{S} = \{\rho_1, \ldots, \rho_N\}$ be the set of $N = \exp(\Omega(d^2))$ states which satisfy the two conditions in Corollary 4.2.6, with corresponding unitary operators $\{U_1, \ldots, U_N\}$. Let $\boldsymbol{x}$ be uniformly random over $[N]$ 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 $\mathcal{K}$) on identical copies of $\rho_{\boldsymbol{x}} = \rho_{\epsilon,U_{\boldsymbol{x}}}$. By Fano's inequality as well as the assumption that the tomography algorithm is accurate to within trace distance $O(\epsilon)$, we know that $I(\boldsymbol{x} : \boldsymbol{y}) \geq \Omega(d^2)$.

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

$$I(\boldsymbol{x} : \boldsymbol{y}) = \sum_{i=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i}) \tag{4.14}$$

where we are using the shorthand $\boldsymbol{y}_{<i}$ to refer to the random variables $\boldsymbol{y}_{i-1}, \ldots, \boldsymbol{y}_1$ as before. For each $i \in [n]$, let $p_{\boldsymbol{y}_i|y_{<i},x}$ be the conditional distribution for the outcome of performing the measurement $\mathcal{M}^{y_{<i}}$ on the state ρ_x, with probabilities given by

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

for each possible outcome y and the set of measurement operators $\{M_y^{y_{<i}}\}_y$ corresponding to the POVM $\mathcal{M}^{y_{<i}}$. Also, define $w^{y_{<i}}$ as the discrete distribution with probabilities $w^{y_{<i}}(y) := \mathbb{E}_{\boldsymbol{U} \sim \mathrm{Haar}} \mathrm{Tr}(M_y^{y_{<i}} \rho_{\epsilon,\boldsymbol{U}})$ for each outcome y. Consider the i^{th} term in the sum in

the right-hand side of (4.14). We may apply our upper bound on the mutual information from Lemma 3.2.1 as well as Definition 4.1.2 for the function $\mathrm{X}(\cdot,\cdot)$ to deduce that

$$\begin{aligned} I(\boldsymbol{x}:\boldsymbol{y}_i|\boldsymbol{y}_{<i}) &= \mathop{\mathbb{E}}_{\boldsymbol{y}'_{<i}\sim p_{\boldsymbol{y}_{<i}}} I(\boldsymbol{x}:\boldsymbol{y}_i|\boldsymbol{y}_{<i}=\boldsymbol{y}'_{<i}) \\ &\leq \mathop{\mathbb{E}}_{\boldsymbol{y}'_{<i}\sim p_{\boldsymbol{y}_{<i}}} \mathop{\mathbb{E}}_{\boldsymbol{x}'\sim p_{\boldsymbol{x}|\boldsymbol{y}_{<i}}} \chi^2(p_{\boldsymbol{y}_i|\boldsymbol{y}'_{<i},\boldsymbol{x}'} \parallel w^{\boldsymbol{y}'_{<i}}) \\ &= \mathop{\mathbb{E}}_{\boldsymbol{y}_{<i}} \mathop{\mathbb{E}}_{\boldsymbol{x}|\boldsymbol{y}_{<i}} \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<i}}, U_{\boldsymbol{x}}) \\ &\leq O\left(\frac{\epsilon^2}{d}\right) + \beta \\ &= O\left(\frac{\epsilon^2(1+\log(m)/d)}{d}\right). \end{aligned} \tag{4.15}$$

The fourth line follows by the assumption that $\mathcal{M}^{y_{<i}} \in \mathcal{K}$ for every $y_{<i}$. Applying this argument to each of the n mutual information terms and making use of the assumption that $m = \exp(O(d))$ we find $n = \Omega\left(d^3/\epsilon^2\right)$ which concludes the proof of Theorem 4.2.1.

Proof of Theorem 4.2.2

To derive the second lower bound in the constant-outcome case we may repeat the above argument using the corresponding tail bound from Lemma 4.1.3. This gives us an $\epsilon/2$-packing of states with corresponding unitaries $U_1, \ldots, U_N$ satisfying $\mathrm{X}(\mathcal{M}, U_i) \leq O(\epsilon^2/d^2) + \beta$ for each $i \in [N]$, implying that the conditional mutual information terms are each $O(\epsilon^2(1 + \log(m))/d^2)$, where we have used the fact that $\beta = O(\epsilon^2 \log(m)/d^2)$. The assumption that $m = \exp(O(d))$ produces the $\Omega(d^4/\epsilon^2(\log(d)))$ lower bound in the constant-outcome case upon invoking Fano's inequality.

4.3 Classical shadows of quantum states

Since classical shadows are believed to have applications in quantum computing, in this section we will focus on the case where the unknown states are of dimension $d = 2^q$, describing a system of q qubits. We will continue to use n to denote the number of samples required for the learning task.

Description of task

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 its 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, a description of the state known as a *classical shadow* [23] can provide an exponential reduction in the number of copies of the state required to learn its relevant properties. More precisely, a classical shadow refers to a successful procedure for the problem described below. Here, *unentangled access* refers to restricting measurements to individual copies of the unknown state, as described in Section 1.4.

Classical shadows problem. Given parameters ϵ, δ, B and unentangled access to n copies of $\rho \in \mathsf{D}(2^q)$ the task is to output a function (classical shadow) $f : \mathsf{Psd}(d)^M \to \mathbb{R}^M$ such that for any fixed collection of M observables $0 \preceq O_1, \ldots, O_M \preceq \mathbb{1}$ satisfying $\mathrm{Tr}(O_i^2) \leq B$ for all $i \in [M]$, it holds that $\|f(O_1, \ldots, O_M) - f_\rho(O_1, \ldots, O_M)\|_\infty \leq \epsilon$ with probability at least $1 - \delta$, where $f_\rho(O_1, \ldots, O_M)_i := \mathrm{Tr}\,(O_i \rho)$ for every $i \in [M]$.

A sample-efficient algorithm

Ref. [23] gives a procedure for computing classical shadows which uses only $n = O(\log(M)B/\epsilon^2)$ efficient, nonadaptive measurements on separate copies of the state ρ. Here, the measurements are implemented by random q-qubit Clifford operators which form a unitary 3-design and then performing a median-of-means estimation of the expectation values. Overall this is an exponential 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 lower bound does not take into account the possibility of adaptive measurements and so we turn to this in Section 4.3.1, considering the worst case where $B = O(d)$.

A worst-case adaptive lower bound

In related work [22] it was shown that in order to predict the expectation values of the 4^q q-qubit Pauli operators with constant-accuracy, one requires $\Omega(2^{q/3})$ copies of the state ρ

	Entangled	Unentangled	Unentangled+Efficient
Upper bound	$\tilde{O}(\log^2(d)\log(M)/\epsilon^4)$ [11]	$O(d\log(M)/\epsilon^2)$ [23]	$O(d\log(M)/\epsilon^2)$ [23]
Lower bound	$\Omega(\log(M)/\epsilon^2)$ [1]	$\Omega(d^{1/3})$ [22]	$\Omega(d\log(M)/\epsilon^2)$ (this work)

Table 4.1: Best known upper and lower bounds for the sample complexity of (ϵ, δ)-shadow tomography of $M \leq \exp(d^2)$ observables under different measurement restriction assumptions. $\tilde{O}$ hides loglog factors in d and $1/\epsilon$.

even if measurements are allowed to be adaptive. This was used to esablish an exponential separation from the case in which all copies of the state can be jointly measured, in which case $O(q)$ copies suffice. Furthermore, since the task of predicting all Pauli expectations reduces to the task of producing classical shadows, this result implies a $\Omega(2^{q/3})$ lower bound on classical shadows in the worst case, meaning that adaptivity cannot lead to an exponential improvement in the copy complexity of this task.

4.3.1 An alternative lower bound with adaptive measurements

In this section, we show how the previous arguments for quantum tomography can be adjusted to give an $n = \Omega(2^q \min\{4^q, \log(M)\}/\epsilon^2)$ lower bound for classical shadows with adaptive measurement protocols that can be efficiently implemented on a quantum computer. Furthermore, the way we obtain this result is by proving this lower bound on the strictly easier task of shadow tomography [1] with unentangled measurements. We therefore find that since there is a matching upper bound using the classical shadows algorithm with random q-qubit Clifford measurements, that approach is optimal not only for the task of classical shadows but also as a procedure for (unentangled) shadow tomography.

(Unentangled) (ϵ, δ)-shadow tomography problem. Given parameters ϵ, δ, unentangled access to n copies of $\rho \in \mathsf{D}(2^q)$, as well as M observables $0 \preceq O_1, \ldots, O_M \preceq \mathbb{1}$, the task is to output a vector $b \in \mathbb{R}^M$ such that with probability at least $1 - \delta$, we have $|b_i - \mathrm{Tr}(O_i\rho)| \leq \epsilon$ for every $i \in [M]$.

Table 4.1 summarizes known results on the sample complexity of shadow tomography under various measurement assumptions. It is clear that when we do not restrict the parameter B from the classical shadows problem (i.e., it can be as large as d) then the output of the classical shadows problem can be used to produce a solution to the shadow tomography problem, setting the parameters ϵ and δ to match. We will prove the following result.

Theorem 4.3.1. *Any procedure for the unentangled shadow tomography problem with the additional restriction that measurements are efficiently implementable must use* $\Omega(2^q \min\{4^q, \log(M)\}/\epsilon^2)$ *copies of the state.*

In Section 4.3.2 we provide a simpler algorithm than the median-of-means protocol of Ref. [23] for optimal unentangled shadow tomography, under the assumption of efficient measurements. The analysis is based on the straightforward approach of "reusing" samples of the unknown state. Taken together, these results address an open question of Aaronson and Rothblum [3] which asked how feasible it might be to perform shadow tomography with unentangled measurements.

Proof of Theorem 4.3.1

Since we are operating under the assumption of measurements implemented using efficient quantum computation, we may fix a finite universal gateset $\mathcal{G}$ with which we implement the measurements. We will show that there exists a collection of L observables $O_1, \ldots, O_L$ whose expectations enable one to uniquely identify a state from $\rho_1, \ldots, \rho_L$, but at the same time whose measurement statistics are uninformative. Moreover, we can take L to be as large as $\exp(\min\{\Omega(d^2), \log(M)\})$ where $d = 2^q$. The lower bound will then follow by using our upper bound on the chi-squared divergence quantity we have been considering in the previous sections on tomography, and combining with Fano's inequality. We first construct our difficult instance of the shadow tomography problem. Let $\boldsymbol{U}$ be a Haar-random unitary taking values in $\mathbb{U}(d)$ and Q be a rank-$d/2$ projector. Lemma 3.1.3 implies that for any fixed rank-$d/2$ projector P, we have

$$\Pr[\mathrm{Tr}(P\boldsymbol{U}Q\boldsymbol{U}^\dagger) \geq d/3] \leq 2\exp(-Cd^2) \tag{4.16}$$

for some universal constant C, by choosing the parameter t in the lemma appropriately. Let $L := \lfloor \frac{1}{4}\exp\left(\min\{Cd^2, \log(M)\}\right)\rfloor$ so that, similar to Proposition 4.2.4 we have the following result: for arbitrary unitary operators $U_1, \ldots, U_L \in \mathbb{U}(d)$ the probability that $\mathrm{Tr}(U_i Q U_i^\dagger \boldsymbol{U} Q \boldsymbol{U}^\dagger) \geq d/3$ for any $i \in [L]$ is at most $1/2$. Let $\mathcal{K} = \{\mathcal{M}_1, \ldots, \mathcal{M}_m\}$ be a set of POVMs from which our measurements are drawn. Then Lemma 4.2.5 applies and we have the following result:

Lemma 4.3.2. *Define* $\mathrm{X}(\mathcal{M}, U)$ *as in Definition 4.1.2 and let* $\beta = \epsilon^2 \ln(6m)/(Cd^2)$*, where* C *is the absolute constant appearing in the tail in Lemma 4.1.3. There exists a set of* L *unitary operators* $U_1, \ldots, U_L \in \mathbb{U}(d)$ *and* Q *a rank-*$d/2$ *orthogonal projector such that*

1. $\mathrm{Tr}(U_i Q U_i^\dagger U_j Q U_j^\dagger) \leq d/3$ *for every* $i, j \in [L], i \neq j$ *and*

2. $\mathrm{X}(\mathcal{M}, U_i) \leq \epsilon^2/d + \beta$ *for every* $i \in [L]$ *and* $\mathcal{M} \in \mathcal{K}$.

Proof. The proof is identical to that for Corollary 4.2.6 except using Eq. (4.16) to impose the first condition using the probabilistic existence argument, rather than Proposition 4.2.4. □

Now, let $\mathcal{S} = \{\rho_1, \ldots, \rho_L\} \subset \mathsf{D}(d)$ be a collection of states of the form

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

for each $i \in [n]$, let $\boldsymbol{x}$ be uniformly random over $[L]$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the measurement outcomes obtained from n unentangled (possibly adaptive) measurements performed on copies of the state $\rho_{\boldsymbol{x}}$. Let $\boldsymbol{x}$ and $\boldsymbol{y}_{<i}$ be identically distributed to $\boldsymbol{x}$ and $\boldsymbol{y}_{<i}$. By chain rule for mutual information, we have

$$\begin{aligned} I(\boldsymbol{x} : \boldsymbol{y}) &= \sum_{i=1}^{n} I(\boldsymbol{x} : \boldsymbol{y}_i | \boldsymbol{y}_{<i}) \\ &\leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\boldsymbol{y}_{<i}} \mathop{\mathbb{E}}_{\boldsymbol{x}|\boldsymbol{y}_{<i}} \mathrm{X}(\mathcal{M}^{\boldsymbol{y}_{<i}}, U_{\boldsymbol{x}}) \\ &= O\left(\frac{n\epsilon^2(1+\log(m)/d)}{d}\right) \end{aligned}$$

where we have omitted some steps since they are identical to those leading to Eq. (4.15). Observing that $m = \exp(O(d))$ by Proposition 4.2.3 we find the upper bound above is at most $O(n\epsilon^2/d)$. On the other hand, it holds that for any $i \in [L]$

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

while by Property 1 in Lemma 4.3.2 we have for any $j \neq i$

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

which means that estimating $\mathrm{Tr}(U_i Q U_i^\dagger \rho_x)$ with $O(\epsilon)$ accuracy for every $i \in [L]$ would allow one to identify the value of $x \in [L]$. Since $L \leq M/4$, a successful protocol for shadow tomography of M observables can do this with some probability at least $1 - \delta$, taking a subset of the input observables to be $U_1 Q U_1^\dagger, \ldots, U_L Q U_L^\dagger$. This in turn implies that $I(\boldsymbol{x} : \boldsymbol{y}) \geq \Omega(\log(L))$ using Fano's inequality, and therefore we arrive at the lower bound of $n = \Omega(d \min\{d^2, \log(M)\}/\epsilon^2)$ as claimed.

4.3.2 Simple algorithm for unentangled shadow tomography

As can be seen from Table 4.1, when there is no restriction placed on the M observables whose expectations are to be learned, the median-of-means protocol due to Ref. [23] is the information-theoretically optimal one for shadow tomography with unentangled, efficient measurements. There, one employs random q-qubit Clifford circuits to make random basis measurements. However, there is an even simpler approach one could take using this same measurement scheme which also leads to optimal performance in terms of sample complexity. Specifically, we will show that just taking straightforward sample means of each observable's expectation reproduces the same upper bound on the overall worst-case sample complexity, which is $n = O(d \log(M)/\epsilon^2)$ where $d = 2^q$.

Suppose our measurements are given by applying a random q-qubit Clifford circuit and then measuring in the standard basis. We may write the set of measurement operators corresponding to the various outcomes as $\{\frac{d}{m}|v_j\rangle\langle v_j|\}_{j=1}^m$ for some $m > 0$ and unit vectors $|v_j\rangle \in \mathbb{C}^d$. The probability that the outcome corresponding to the j^{th} measurement operator is obtained is $\frac{d}{m}\langle v_j|\rho|v_j\rangle$. Define $\hat{\rho}(v) = (d+1)|v\rangle\langle v| - \mathbb{1}$, and let $\boldsymbol{v}$ be the random variable corresponding to the outcome that is obtained. As was demonstrated in Section 1.5, using the defining property of t-designs one may verify that $\mathbb{E}\,\hat{\rho}(\boldsymbol{v}) = \rho$. We also require the following property, which holds since our measurement operators form a 3-design.

Proposition 4.3.3 (Prop. S1, Sec. 5 in Suppl. Materials of Ref. [23])**.** *Let $0 \preceq X \preceq \mathbb{1}$ be a positive semidefinite operator acting on $\mathbb{C}^d$ and $\hat{\rho}(\cdot)$, $\boldsymbol{v}$ be as defined above. Then*

$$\mathrm{Var}\left[\mathrm{Tr}(X\hat{\rho}(\boldsymbol{v}))\right] \leq 3\mathrm{Tr}(X^2).$$

Now suppose that the observables of interest are $0 \preceq O_1, \ldots, O_M \preceq \mathbb{1}$ satisfying $\mathrm{Tr}(O_i^2) \leq B\ \forall i \in [M]$, and define the function $f_i(v) := \mathrm{Tr}(O_i\hat{\rho}(v))$ for any observable O_i and outcome v. Then $f_i(\boldsymbol{v})$ with mean $\mathbb{E}\, f_i(\boldsymbol{v}) = \mathrm{Tr}(O_i\,\mathbb{E}\,\hat{\rho}(\boldsymbol{v})) = \mathrm{Tr}(O_i\rho)$ is an unbiased estimator for the i^{th} quantity we wish to estimate, $\mathrm{Tr}(O_i\rho)$. Suppose one performs the measurement n times and obtains the i.i.d. outcome random variables $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Let us consider the empirical mean of the i^{th} estimator f_i. By Bernstein's inequality (see Appendix B.3), we have for any $\epsilon > 0$ that

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

where $\sigma^2 := \frac{1}{n}\sum_{j=1}^n \mathrm{Var}[f_i(\boldsymbol{v}_j)]$ and K is such that $|f_i(\boldsymbol{v}_j) - \mathrm{Tr}(O_i\rho)| \leq K$ with probability 1 for all $j \in [n]$. Now observe that by definition $\|f_i\|_\infty \leq d+1$ so K can be taken to be $O(d)$,

and by Proposition 4.3.3 $\sigma^2 \leq B \leq d$. This implies that, taking n to be $O(d \log(M/\delta)/\epsilon^2)$ we obtain that the probability above is at most δ/M. By union bound, we can estimate $\mathbb{E}\, f_i(\boldsymbol{v}) = \mathrm{Tr}(O_i\rho)$ for all $i \in [M]$ to additive error ϵ using just these $n = O(d \log(M/\delta)/\epsilon^2)$ samples of the state ρ, with failure probability at most δ.

4.4 Open problems

Can we use similar techniques to obtain a lower bound which takes into account the maximum Frobenius norm of the observables of interest? (This is the parameter B in the classical shadows problem.) In Theorem 4.2.2 we incur a $\log(d)$ factor in the denominator of the lower bound for tomography with constant-outcome measurements. Can this be improved? Note that this factor also appears in the denominator of the lower bound due to Flammia et al. [16]. Is there a way to incorporate rank-dependence into these lower bounds? This seems to be related to the B-dependent lower bound for classical shadows.

References

[1] Scott Aaronson. "Shadow tomography of quantum states". In: *Proceedings of the 50th Annual ACM Symposium on Theory of Computing*. ACM, June 2018, pp. 325–338. DOI: 10.1145/3188745.3188802. URL: https://doi.org/10.1145/3188745.3188802.

[2] Scott Aaronson and Daniel Gottesman. "Improved simulation of stabilizer circuits". In: *Physical Review A* 70.5 (2004). ISSN: 1094-1622. DOI: 10.1103/physreva.70.052328. URL: http://dx.doi.org/10.1103/PhysRevA.70.052328.

[3] Scott Aaronson and Guy N. Rothblum. *Gentle Measurement of Quantum States and Differential Privacy*. 2019. arXiv: 1904.08747 [quant-ph].

[4] Scott Aaronson et al. "Online learning of quantum states". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124019. ISSN: 1742-5468. DOI: 10.1088/1742-5468/ab3988. URL: http://dx.doi.org/10.1088/1742-5468/ab3988.

[5] Andris Ambainis and Joseph Emerson. "Quantum t-designs: t-wise Independence in the Quantum World". In: *2007 22nd Annual IEEE Conference on Computational Complexity*. Los Alamitos, CA, USA: IEEE Computer Society, 2007, pp. 129–140. DOI: 10.1109/CCC.2007.26. URL: https://doi.ieeecomputersociety.org/10.1109/CCC.2007.26.

[6] George B. Arfken, Hans J. Weber, and Donald Spector. "Mathematical Methods for Physicists, 4th ed." In: *American Journal of Physics* 67.2 (1999), pp. 165–169. DOI: 10.1119/1.19217. URL: https://doi.org/10.1119/1.19217.

[7] Sanjeev Arora, Elad Hazan, and Satyen Kale. "The Multiplicative Weights Update Method: a Meta-Algorithm and Applications". In: *Theory of Computing* 8.6 (2012), pp. 121–164. DOI: 10.4086/toc.2012.v008a006. URL: http://www.theoryofcomputing.org/articles/v008a006.

[8] Sanjeev Arora and Satyen Kale. "A combinatorial, primal-dual approach to semidefinite programs". In: *Proceedings of the 39th annual ACM symposium on Theory of computing*. ACM Press, 2007. DOI: 10.1145/1250790.1250823. URL: https://doi.org/10.1145/1250790.1250823.

[9] Sanjeev Arora and Satyen Kale. "A Combinatorial, Primal-Dual Approach to Semidefinite Programs". In: *Journal of the ACM* 63.2 (May 2016), pp. 1–35. DOI: 10.1145/2837020. URL: https://doi.org/10.1145/2837020.

[10] Srinivasan Arunachalam, Yihui Quek, and John Smolin. *Private learning implies quantum stability*. 2021. arXiv: 2102.07171 [quant-ph].

[11] Costin Bădescu and Ryan O'Donnell. *Improved quantum data analysis*. 2020. arXiv: 2011.10908 [quant-ph].

[12] Fernando G. S. L. Brandão, Richard Kueng, and Daniel Stilck França. *Fast and robust quantum state tomography from few basis measurements*. 2020. arXiv: 2009.08216 [quant-ph].

[13] Sebastien Bubeck, Sitan Chen, and Jerry Li. *Entanglement is Necessary for Optimal Quantum Property Testing*. 2020. arXiv: 2004.07869 [quant-ph].

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Apr. 2005. DOI: 10.1002/047174882x. URL: https://doi.org/10.1002/047174882x.

[15] Robert M Fano. *Transmission of information: a statistical theory of communications*. MIT Press, 1966.

[16] Steven T. Flammia et al. "Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators". In: *New Journal of Physics* 14.9 (Sept. 2012), p. 095022. ISSN: 1367-2630. DOI: 10.1088/1367-2630/14/9/095022. URL: http://dx.doi.org/10.1088/1367-2630/14/9/095022.

[17] Sidney Golden. "Lower Bounds for the Helmholtz Function". In: *Physical Review* 137 (4B 1965), B1127–B1128. DOI: 10.1103/PhysRev.137.B1127. URL: https://link.aps.org/doi/10.1103/PhysRev.137.B1127.

[18] Jeongwan Haah et al. "Sample-optimal tomography of quantum states". In: *IEEE Transactions on Information Theory* (2017), pp. 5628 –5641. ISSN: 1557-9654. DOI: 10.1109/tit.2017.2719044. URL: http://dx.doi.org/10.1109/TIT.2017.2719044.

[19] Patrick Hayden, Debbie W. Leung, and Andreas Winter. "Aspects of Generic Entanglement". In: *Communications in Mathematical Physics* 265.1 (Mar. 2006), pp. 95–117. DOI: 10.1007/s00220-006-1535-6. URL: https://doi.org/10.1007/s00220-006-1535-6.

[20] Elad Hazan. "Introduction to Online Convex Optimization". In: *Foundations and Trends in Optimization* 2.3-4 (2016), pp. 157–325. ISSN: 2167-3888. DOI: 10.1561/2400000013. URL: http://dx.doi.org/10.1561/2400000013.

[21] Jean-Baptiste Hiriart-Urruty and Lemaréchal Claude. "Fundamentals of convex analysis". In: Springer, 2004.

[22] Hsin-Yuan Huang, Richard Kueng, and John Preskill. *Information-theoretic bounds on quantum advantage in machine learning*. 2021. arXiv: 2101.02464 [quant-ph].

[23] Hsin-Yuan Huang, Richard Kueng, and John Preskill. "Predicting many properties of a quantum system from very few measurements". In: *Nature Physics* 16.10 (June 2020), pp. 1050–1057. DOI: 10.1038/s41567-020-0932-7. URL: https://doi.org/10.1038/s41567-020-0932-7.

[24] F. Huszár and N. M. T. Houlsby. "Adaptive Bayesian Quantum Tomography". In: *Physical Review A* 85 (5 2012), p. 052120. DOI: 10.1103/PhysRevA.85.052120. URL: https://link.aps.org/doi/10.1103/PhysRevA.85.052120.

[25] R. Z. Khas'minskii. "A Lower Bound on the Risks of Non-Parametric Estimates of Densities in the Uniform Metric". In: *Theory of Probability & Its Applications* 23.4 (Sept. 1979), pp. 794–798. DOI: 10.1137/1123095. URL: https://doi.org/10.1137/1123095.

[26] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. "Low rank matrix recovery from rank one measurements". In: *Applied and Computational Harmonic Analysis* 42.1 (2017), pp. 88–116. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2015.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1063520315001037.

[27] D. H. Mahler et al. "Adaptive Quantum State Tomography Improves Accuracy Quadratically". In: *Physical Review Letters* 111 (18 2013), p. 183601. DOI: 10.1103/PhysRevLett.111.183601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.111.183601.

[28] Elizabeth S. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Vol. 218. Cambridge Tracts in Mathematics. Cambridge University Press, July 2019. DOI: 10.1017/9781108303453.

[29] Ashley Montanaro. *Pretty simple bounds on quantum state discrimination*. 2019. arXiv: 1908.08312 [quant-ph].

[30] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511976667.

[31] Ryan O'Donnell and John Wright. *Efficient quantum tomography*. 2015. arXiv: 1508.01907 [quant-ph].

[32] Igal Sason and Sergio Verdú. "f-Divergence Inequalities". In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 5973–6006. DOI: 10.1109/TIT.2016.2603151.

[33] G.I. Struchalin et al. "Experimental Estimation of Quantum State Properties from Classical Shadows". In: *PRX Quantum* 2 (1 2021), p. 010307. DOI: 10.1103/PRXQuantum.2.010307. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.010307.

[34] Colin J. Thompson. "Inequality with Applications in Statistical Mechanics". In: *Journal of Mathematical Physics* 6.11 (Nov. 1965), pp. 1812–1813. DOI: 10.1063/1.1704727. URL: https://doi.org/10.1063/1.1704727.

[35] Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth. "Matrix Exponentiated Gradient Updates for On-line Learning and Bregman Projection". In: *Journal of Machine Learning Research* 6.34 (2005), pp. 995–1018. URL: http://jmlr.org/papers/v6/tsuda05a.html.

[36] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. New York, NY: Springer New York, 2009, pp. 77–135. ISBN: 978-0-387-79052-7. DOI: 10.1007/978-0-387-79052-7_2. URL: https://doi.org/10.1007/978-0-387-79052-7_2.

[37] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

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

[39] John Wright. "How to learn a quantum state". PhD thesis. Carnegie Mellon University, 2016.

[40] Akram Youssry, Christopher Ferrie, and Marco Tomamichel. "Efficient online quantum state estimation using a matrix-exponentiated gradient method". In: *New Journal of Physics* 21.3 (2019), p. 033006. ISSN: 1367-2630. DOI: 10.1088/1367-2630/ab0438. URL: http://dx.doi.org/10.1088/1367-2630/ab0438.

[41] Nengkun Yu. *Sample efficient tomography via Pauli Measurements*. 2020. arXiv: 2009.04610 [quant-ph].

APPENDICES

Appendix A

Haar integrals

The Haar measure μ is the unique unitarily invariant probability measure on the space of unitary operators, $\mathbb{U}(d)$. 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

$$\Phi_k(X) = \int_{\mathbb{U}(d)} U^{\otimes k} X (U^\dagger)^{\otimes k} \mathrm{d}\mu(U) \tag{A.1}$$

which are referred to as "twirl" operations. In the rest of this section, we will be concerned with evaluating 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 [38], we will make use of an important result on the structure of permutation-invariant operators.

Theorem A.0.1 (Theorem 7.15 in [38])**.** *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. $[X, U^{\otimes k}] = 0 \ \forall U \in \mathbb{U}(d)$.
2. $X = \sum_{\pi\in S_k} v(\pi) W_\pi$ *for some choice of* $v \in \mathbb{C}^{|S_k|}$.

Here, W_π is the permutation operator corresponding to the permutation $\pi \in S_k$, and S_k is the symmetric group on $\{1, \ldots, k\}$.

(See Section 1.2 for more on the permutation operator.) Since $\Phi_k(X)$ satisfies the first condition, we can apply the theorem to write it as a linear combination of permutation operators. This will help us evaluate the Haar integrals which arise in the remainder of the proof.

Proposition A.0.2. *Let $d > 1$ be a positive integer, Q a rank-r projector with $r < d$, and $\boldsymbol{U}$ a Haar-random unitary operator. It holds that*

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

and

$$\mathbb{E}(\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes 2} = \frac{r}{d(d^2-1)}\left[(rd-1)\mathbb{1} + (d-r)W\right]$$

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

Proof. We can write the expectations as

$$\int_{\mathbb{U}(d)} (UQU^\dagger)^{\otimes k} \mathrm{d}\mu(U) = \Phi_k(Q^{\otimes k})$$

for $k = 1, 2$ respectively. By Theorem A.0.1, each of these must be equal to a linear combination of permutation operators. For $k = 1$, 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 projector, taking the trace of both sides and solving for κ yields $\kappa = r/d$. For $k = 2$ we have

$$\mathbb{E}(\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes 2} = \alpha\mathbb{1}\otimes\mathbb{1} + \beta W \tag{A.2}$$

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

$$\mathrm{Tr}(Q)^2 = r^2 = \alpha d^2 + \beta d, \quad \mathrm{Tr}(Q^2) = r = \alpha d + \beta d^2.$$

This allows us to solve for α, β:

$$\alpha = \frac{r(rd-1)}{d(d^2-1)}, \quad \beta = \frac{r(d-r)}{d(d^2-1)}. \tag{A.3}$$

This concludes the proof of the proposition. □

Proof of Lemma 2.2.3

We are interested in evaluating quantities of the form

$$\mathbb{E}\,\mathrm{Tr}(\boldsymbol{U}Q\boldsymbol{U}^\dagger(\omega-\rho))^k = \mathrm{Tr}\left(\mathbb{E}(\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes k}(\omega-\rho)^{\otimes k}\right) \tag{A.4}$$

for $k = 2, 4$, where $\boldsymbol{U}$ comprises a unitary 4-design, Q is a rank-$d/2$ orthogonal projector, ω and ρ are quantum states, and the equality follows from linearity of the trace function. Using the fact that the set of unitaries forms a 4-design, we can write the expectation on the right-hand side for either $k = 2$ or $k = 4$ as

$$\mathbb{E}(\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes k} = \int_{\mathbb{U}(d)} (UQU^\dagger)^{\otimes k}\mathrm{d}\mu(U) = \Phi_k(Q^{\otimes k})$$

which, by Theorem A.0.1, must be equal to a linear combination of permutation operators. We have already seen in Proposition A.0.2 that for the case where $k = 2$ and the rank of Q is $d/2$ we obtain

$$\mathbb{E}(\boldsymbol{U}Q\boldsymbol{U}^\dagger)^{\otimes 2} = \alpha\mathbb{1}\otimes\mathbb{1} + \beta W \tag{A.5}$$

where W is the swap operator and

$$\alpha = \frac{d^2-2}{4(d^2-1)}, \quad \beta = \frac{d}{4(d^2-1)}. \tag{A.6}$$

Furthermore, substituting (A.5) into the right-hand side of (A.4) when $k = 2$ and making use of the fact that $\mathrm{Tr}(\omega-\rho) = 0$ gives us

$$\mathbb{E}\mathrm{Tr}(\boldsymbol{U}Q\boldsymbol{U}^\dagger(\omega-\rho))^2 = \beta\mathrm{Tr}(W(\omega-\rho)^{\otimes 2}) = \frac{d}{4(d^2-1)}\left\|\omega-\rho\right\|_{\mathrm{F}}^2$$

where the final equality made use of (A.6) along with the identity $\mathrm{Tr}(W(\omega-\rho)^{\otimes 2}) = \|\omega-\rho\|_{\mathrm{F}}^2$. Noting that $\frac{d}{4(d^2-1)} = \Theta(1/d)$ completes the proof of the first equation in Lemma 2.2.3.

Before continuing with the second part of the proof, we arm ourselves with a generalization of the above reasoning which we can use when $k = 4$.

Proposition A.0.3. *Let $s(\pi)$ be the number of cycles in the permutation π, and $S_{\mathrm{even}} \subset S_4$ be the set of all permutations in S_4 having either two 2-cycles or one 4-cycle. Then*

$$\int_{\mathbb{U}(d)} \mathrm{Tr}(UQU^\dagger(\omega-\rho))^4\mathrm{d}\mu(U) \leq \sum_{\pi\in S_{\mathrm{even}}} |v(\pi)|\left\|\omega-\rho\right\|_{\mathrm{F}}^4$$

where $v \in \mathbb{R}^{|S_4|}$ *satisfies*

$$\sum_{\sigma \in S_4} v(\sigma) d^{s(\pi\sigma)} = \left(\frac{d}{2}\right)^{s(\pi)}$$

for every $\pi \in S_4$.

Proof. To prove the first relation, we use the fact that

$$\int_{\mathbb{U}(d)} \mathrm{Tr}(UQU^\dagger(\rho - \sigma))^4 \mathrm{d}\mu(U) = \mathrm{Tr}(\Phi_4(Q^{\otimes 4})(\rho - \sigma)^{\otimes 4}). \tag{A.7}$$

Since, by Theorem A.0.1

$$\Phi_4(Q^{\otimes 4}) = \sum_{\pi \in S_4} v(\pi) W_\pi \tag{A.8}$$

for some $v \in \mathbb{R}^{|S_4|}$ (the fact that v is real comes from the proof of the second relation), we have that the right-hand side of (A.7) is

$$\sum_{\pi \in S_4} v(\pi) \mathrm{Tr}(W_\pi(\rho - \sigma)^{\otimes 4}). \tag{A.9}$$

Therefore, we must evaluate $\mathrm{Tr}(W_\pi(\rho - \sigma)^{\otimes 4})$ for the various permutations in S_4, which is facilitated by the following observation:

Proposition A.0.4. *Let* $W_\pi \in (\mathbb{C}^{d\times d})^{\otimes 4}$ *be the permutation operator for* $\pi \in S_4$ *and* $A_i \in \mathbb{C}^{d\times d}$ *be Hermitian operators for* $i = 1, \ldots, 4$. *It holds that*

$$\mathrm{Tr}(W_\pi(A_1 \otimes A_2 \otimes A_3 \otimes A_4)) = \prod_{\mathrm{cyc} \in \mathrm{cycles}(\pi)} \mathrm{Tr}\left(\left(\prod_{i \in \mathrm{cyc}} A_i\right)^\dagger\right).$$

For example, if $\pi = (2)(134)$, we will obtain

$$\mathrm{Tr}\left(W_\pi(\omega - \rho)^{\otimes 4}\right) = \mathrm{Tr}\,(\omega - \rho)\,\mathrm{Tr}\left((\omega - \rho)^3\right) = 0$$

because $\mathrm{Tr}(\omega - \rho) = 0$. By this reasoning, any term on the right-hand side of (A.9) corresponding to a permutation with a 1-cycle will evaluate to zero, so it remains to upper bound $v(\pi)\mathrm{Tr}(W_\pi(\omega - \rho)^{\otimes 4})$ whenever π has two 2-cycles or one 4-cycle. In the first case we

obtain $v(\pi)\,\|\omega-\rho\|_{\mathrm{F}}^4$, while in the second we get $v(\pi)\,\|\omega-\rho\|_4^4 \le |v(\pi)|\,\|\omega-\rho\|_{\mathrm{F}}^4$, which concludes the proof of the inequality.

One may obtain the second relation in the proposition by left-multiplying each side of (A.8) by W_π and taking the trace, for each $\pi \in S_4$. Then, by once again invoking Observation A.0.4 and making use of the fact that $\mathrm{Tr}(Q^k) = d/2$ for any positive integer k, the desired set of equalities holds. □

We have already seen the derivation of the second moment in Lemma 2.2.3, so it remains to show the claimed inequality

$$\mathbb{E}\mathrm{Tr}\left(\boldsymbol{U}Q\boldsymbol{U}^\dagger(\omega-\rho)\right)^4 \le O\left(d^{-2}\right)\|\omega-\rho\|_{\mathrm{F}}^4\,. \tag{A.10}$$

By the first relation in Proposition A.0.3 it suffices to show that $|v(\pi)| = O(d^{-2})$ for every $\pi \in S_4$ with π having either two 2-cycles or one 4-cycle i.e., $\pi \in S_{\mathrm{even}}$. Moreover, the second relation gives a system of linear equations with a unique solution $v \in \mathbb{R}^{|S_4|}$ satisfying the requirement for each $v(\pi)$. In matrix form, this system of equations is (for some ordering of the elements in S_4)

$$\begin{pmatrix} d^4 & d^3 & d^3 & d^2 & \dots \\ d^3 & d^4 & d^2 & d^3 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ d^2 & d & d^3 & d^2 & \dots \end{pmatrix}_{24\times 24} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ \vdots \end{pmatrix}_{24\times 1} = \begin{pmatrix} d^4/16 \\ d^3/8 \\ \vdots \\ d^2/4 \end{pmatrix}_{24\times 1}$$

yielding, for example,

$$v_{24} = \frac{d^2-6}{16(d^4-10d^2-9)} = O(d^{-2}),$$

and similarly for each element v_i corresponding to a permutation $\pi \in S_{\mathrm{even}}$. The full calculation is omitted here. We remark that there is in all likelihood an easier way to perform this integration, but the method we use is consistent with how we evaluated the other Haar integrals appearing in this thesis.

Appendix B

Miscellaneous facts

B.1 Incomplete Gamma function

The (upper) incomplete Gamma function $\Gamma(\cdot, \cdot)$ is defined by the integral

$$\Gamma(p, a) := \int_a^\infty t^{p-1} e^{-t} \mathrm{d}t = (p-1)!\, e^{-a} \sum_{k=0}^{p-1} \frac{a^k}{k!}$$

where the series representation is valid for p a positive integer (see for example Chapter 10 in Ref. [6]). Factoring, we find that the series is equal to

$$e^{-a} a^{p-1} \left(1 + (p-1)! \sum_{k=0}^{p-2} \frac{a^{k-p+1}}{k!} \right) = e^{-a} a^{p-1} \left(1 + \frac{p-1}{a} + \frac{(p-1)(p-2)}{a^2} + \cdots + \frac{(p-1)!}{a^{p-1}} \right)$$

which is $e^{-a} a^{p-1}(1 + o(1))$ whenever $p = o(a)$, as required in the proof of Lemma 4.1.4.

B.2 Some matrix inequalities

First, we state without proof the standard result of Golden [17] and Thompson [34].

Proposition B.2.1 (Golden-Thompson inequality)**.** *For any symmetric matrices $A, B \in \mathbb{R}^{d \times d}$ we have*

$$\mathrm{Tr}(\exp(A + B)) \leq \mathrm{Tr}(\exp(A) \exp(B)).$$

We also have the following relation.

Proposition B.2.2. *Let $A \in \mathbb{R}^{d\times d}$ be a Hermitian matrix with $\|A\| \leq 1$. It holds that*

$$\exp(-A) \preceq \mathbb{1} - A + A^2.$$

Proof. Firstly, it holds that $\exp(-x) \leq 1 - x + x^2$ for all $|x| \leq 1$. Let $PDP^\intercal$ be the eigendcomposition of the Hermitian matrix A, for some diagonal matrix D and orthogonal matrix P. It suffices to show that the claimed inequality holds for D, since if $\exp(-D) \preceq M$ then $P\exp(-D)P^\intercal = \exp(-PDP^\intercal) \preceq M$ as well, and $\exp(-D) \preceq \mathbb{1} - D + D^2$ follows from the upper bound in the real number case. □

B.3 Bernstein's inequality

Here, we state without proof a version of Bernstein's inequality that applies to random variables with bounded distributions. This concentration inequality can be found in Ref. [37], for example.

Theorem B.3.1 (Theorem 2.8.4 in Ref. [37])**.** *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 $\sigma^2 := \sum_{i=1}^{n} \mathbb{E}\,\boldsymbol{X}_i^2$.