Principal eigenstate classical shadows

Daniel Grier DGRIER@UCSD.EDU
Department of Mathematics and Department of Computer Science and Engineering, University of California, San Diego

Hakop Pashayan HAKOPSADVENTURES@GMAIL.COM
Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Germany

Luke Schaeffer LRSCHAEF@UMD.EDU
Joint Center for Quantum Information and Computer Science, University of Maryland, College Park

Editors: Shipra Agrawal and Aaron Roth

Abstract

Given many copies of an unknown quantum state ρ, we consider the task of learning a classical description of its principal eigenstate. Namely, assuming that ρ has an eigenstate $|\phi\rangle$ with (unknown) eigenvalue $\lambda > 1/2$, the goal is to learn a (classical shadows style) classical description of $|\phi\rangle$ which can later be used to estimate expectation values $\langle\phi|O|\phi\rangle$ for any O in some class of observables. We consider the sample-complexity setting in which generating a copy of ρ is expensive, but joint measurements on many copies of the state are possible. We present a protocol for this task scaling with the principal eigenvalue λ and show that it is optimal within a space of natural approaches, e.g., applying quantum state purification followed by a single-copy classical shadows scheme. Furthermore, when λ is sufficiently close to 1, the performance of our algorithm is optimal—matching the sample complexity for pure state classical shadows.

Keywords: List of keywords

1. Introduction

A key principle of algorithm design is to never do more work than is needed for the task at hand. Consider the problem of identifying some unknown quantum state ρ by measuring several copies of it. It has long been known that obtaining a complete description of such a state (say, by producing an estimate $\hat{\rho}$ close in trace distance) requires a number of copies which grows linearly (or more) with the dimension of the Hilbert space. Such a strong requirement on the number of copies makes it nearly impossible to experimentally realize such tomographic protocols on all but the smallest quantum systems.

Fortunately, a complete description of ρ is unnecessary for many applications, allowing for dramatically simpler estimation protocols. Suppose, for example, you wish to estimate the fidelity of a state ρ produced by an experimental quantum device with some target pure state, say, to benchmark your device. In this case, the number of copies you must prepare scales only with your desired precision, not with the dimension of the ambient space, making the entire procedure much more practical.

The fidelity estimation protocol is a special case of a recent and enormously popular framework for predicting properties of unknown quantum states called *classical shadows* introduced by Huang et al. (2020). In this setting, many copies of the unknown state ρ are

measured and condensed into a classical bit string. This classical description can later be used to estimate $\mathrm{Tr}(O\rho)$ for any O in some class of Hermitian observables with very high probability. The success of the classical shadows framework (Cerezo et al., 2021; Bharti et al., 2022) motivates a deeper consideration into how it can be further improved to model practical quantum learning tasks as well as how it can be made more sample efficient.

In many practical scenarios it is not properties of the state ρ that one wants to learn, but rather those of its top eigenstate. A natural setting where principal eigenstates become the focal object is when one only has access to noisy copies of a target state $|\phi\rangle$. In the case of global depolarizing noise acting on a d-dimensional Hilbert space, the noisy state is $\rho = (1-\eta)|\phi\rangle\langle\phi| + \eta(I - |\phi\rangle\langle\phi|)/(d-1)$. For $\eta < 1/2$, the principal eigenstate remains $|\phi\rangle$. Hence, in the case of global depolarizing noise, the target state can be recovered from the principal eigenstate. In fact, this remains true for other practical noise models as well (Koczor, 2021a; Huggins et al., 2021). Furthermore, settings where one only has access to noisy copies are natural. For example, consider a scenario where copies of a noisy quantum state ρ are prepared by a quantum sensor operating in non-ideal environmental conditions and fed into a powerful quantum processor to extract data. Indeed Ref. Yamamoto et al. (2022) considers such a scenario in the setting of unknown fluctuating noise. Other proposed applications in related work include noise suppression for noisy intermediate-scale quantum computation Koczor (2021b); Huggins et al. (2021); Czarnik et al. (2021); Seif et al. (2023); Zhou and Liu (2022).

In this paper, we ask what happens when you combine classical shadows with principal eigenstate estimation. Namely, what is the complexity of estimating observable expectation values with respect to the dominant eigenvector of ρ rather than ρ itself? To this end, we introduce the following "principal eigenstate classical shadows" task:

Principal eigenstate classical shadows

Learning	*Input:*	Copies of $\rho = (1-\eta)\phi + \eta\sigma$ with principal eigenstate $\phi = \|\phi\rangle\langle\phi\|$, $\eta < 1/2$, and $\mathrm{Tr}(\phi\sigma) = 0$
	Output:	Classical description $\hat{\phi}$
Estimation	*Input:*	Hermitian observable O with $\|O\|_\infty \leq 1$ and classical description $\hat{\phi}$
	Output:	Compute E such that $\|\langle\phi\|O|\phi\rangle - E| \leq \epsilon$

In this work, we focus on the goal of solving the principal eigenstate classical shadows problem with the fewest copies of the input state ρ. That is, we want to determine the *sample complexity* of this task since producing a copy of ρ is usually considered to be a resource-intensive task.

One of the key parameters of this task is η—the *principal deviation*—which determines how far ρ is from its principal eigenstate. Notice that it is this deviation that prevents traditional classical shadow approaches from achieving a high degree of accuracy on this task. That is, classical shadows protocols for the state ρ can only be accurate up to additive error η on the state $|\phi\rangle$ since $|\mathrm{Tr}(O\rho) - \langle\phi|O|\phi\rangle| = \eta$ for $O = |\phi\rangle\langle\phi|$.

Nevertheless, even if you were satisfied with an estimate to accuracy η—a setting in which you could theoretically still use traditional classical shadows approaches—there is

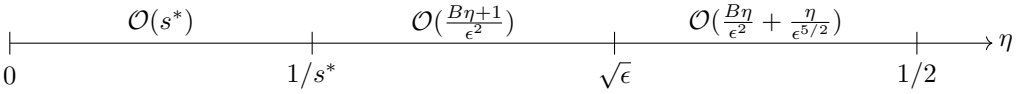
still reason to suspect that more sample-efficient algorithms exist. Intuitively, traditional shadow estimation algorithms do not take advantage of the purity of the underlying state $|\phi\rangle$ that we wish to measure. Indeed, in the noiseless setting (i.e., $\eta = 0$), any shadow algorithm which does not take advantage of the purity of the underlying state is provably suboptimal (Grier et al., 2022). Furthermore, the sample complexity may improve by using joint measurements—i.e., measurements on multiple copies of ρ simultaneously (i.e., $\rho^{\otimes n}$ for some n) using a single entangling positive operator-valued measure (POVM).

We will show that this intuition is correct. That is, we give algorithms for the principal eigenstate classical shadows problem that leverage these insights to be significantly more sample-efficient than classical shadow protocols for generic states. One might quite naturally wonder if our algorithm is simply the combination of other powerful subroutines for processing quantum states, of which many are possibly relevant (Cirac et al., 1999; Keyl and Werner, 2001; Fu, 2016; Childs et al., 2023; O'Donnell and Wright, 2016). For example, it is true that we could design an algorithm for principal eigenstate classical shadows by first de-noising (sometimes referred to as "purifying") the state ρ into the state $|\phi\rangle$ (Cirac et al., 1999; Keyl and Werner, 2001; Lloyd et al., 2014; Fu, 2016; Childs et al., 2023), and then applying a classical shadows algorithm (Huang et al., 2020). We show that the sample complexity of this approach is *worse* than the robust algorithm of this paper which solves the principal eigenstate shadows problem directly. In other words, even though both of the subroutines mentioned above are optimal for their respective subtasks, they are nevertheless doing more work than needed when combined to perform observable estimation on the principal eigenstate.

1.1. Main result

Our classical shadows protocol is actually a suite of algorithms that depend on the principal deviation η of the underlying state $\rho = (1-\eta)\phi + \eta\sigma$. We do not require a priori knowledge of η. We will see that there are three η regimes and as η decreases from $1/2$ to zero, sample complexity also decreases reaching a minimum at the optimal sample complexity for learning pure states (Grier et al., 2022). Surprisingly, this occurs before η becomes zero. These sample complexities are given in the following theorem:

Theorem 1 *There exists a protocol (comprised of separate learning and estimation algorithms) for solving the principal eigenstate classical shadows task with high probability that has three regimes of sample complexity determined by the deviation η shown below*



where $B \geq \mathrm{Tr}(O^2)$ is the squared-Frobenius norm of observable O and $s^* := \frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}$ is the optimal sample complexity for classical shadows on pure states.[1] *Furthermore, in all*

1. Technically, solely in the $\eta \in (\sqrt{\epsilon}, 1/2)$ regime, we invoke a purification procedure of Fu (2016); Childs et al. (2023) that only works on depolarized states, i.e., $\rho = (1-\eta)\phi + \eta\sigma$ for $\sigma = (I-\phi)/(d-1)$. However, based on ongoing/unpublished work, we claim that this purification procedure can be generalized to allow for arbitrary σ.

regimes, the protocol incurs a $\mathcal{O}(\log M)$ factor in sample complexity to approximate M observable expectation values (all to ϵ accuracy) with high probability.

These bounds may at first seem somewhat arbitrary, so let's spend a few moments to put them in context. First, it is worth noting that the optimal sample complexity in the zero-deviation setting is $\tilde{\Theta}(\epsilon^{-1}\sqrt{B} + \epsilon^{-2})$ as determined by Grier et al. (2022). In other words, in the first regime where the deviation is quite small (i.e., $\eta \leq 1/s^*$), the sample complexity is identical to that of the optimal zero-deviation measurement protocol. Moreover, our measurement protocol in this regime is actually the same measurement procedure as used for zero deviation. However, this is not to say that the analysis of this protocol is trivial or in any sense a black box reduction to the pure case. In fact, the bulk of the technical work in this paper is spent addressing this setting.

The sample complexities in the remaining two settings are shown in some sense by a reduction to the first setting. In the second regime, we measure independent blocks of the unknown state, and post-process these measurement results with a simple averaging procedure. In the third regime, when the noise is the most extreme, we must first pre-process the input by an explicit purification protocol that uses multiple copies of ρ to distill a new quantum state with smaller deviation. We then proceed by invoking the procedure above. For this procedure, we determine the optimal choice of the number of copies to be jointly measured, averaged, and purified. We also present an estimation protocol for η that can be used to achieve an overall sample complexity which, up to big-$\mathcal{O}$ notation, matches that of our procedure when using an optimal choice of parameters and a *known* η.

Finally, we note that we can always employ the standard median-of-means trick (Lugosi and Mendelson, 2019; Lerasle, 2019; Huang et al., 2020) to amplify the success probability of obtaining an accurate estimate. In this way, to estimate M distinct observables $\{O_i \mid \mathrm{Tr}(O_i^2) \leq B\}_{i=1}^M$, we incur a factor of $\log(M)$ in the sample complexity, as is typical with shadow estimation protocols (Huang et al., 2020).

1.2. Technical challenges

One of the central tools used in tomographic protocols for pure states (Massar and Popescu, 1995; Wright, 2015; Grier et al., 2022) is a continuous POVM proportional to $\{|\psi\rangle\langle\psi|^{\otimes n}\}_\psi$ that we call the *standard symmetric joint measurement* (see Definition 5). Intuitively, this POVM is more likely to output a measurement outcome ψ the closer $|\psi\rangle$ is to the measured unknown state ρ. Not only is this measurement optimal for pure state tomography (Massar and Popescu, 1995), but it is also surprisingly easy to analyze in many cases due to its tight connection with representation theory and the symmetric subspace (Harrow, 2013). Indeed, it is this connection that allows for a simple analysis of the original classical protocol (Huang et al., 2020; Mele, 2023).

The main conceptual bottleneck for our analysis is that the unknown state ρ may have small overlap with the symmetric subspace, rendering the standard symmetric joint measurement ineffective. Of course, there are more powerful representation-theoretic tools for learning mixed states, but these tend to incur a factor of the Hilbert space dimension, which is prohibitively large for many applications of classical shadows (Haah et al., 2016; O'Donnell and Wright, 2016). That said, when the deviation η is small, ρ will still be close

enough to the symmetric subspace that the standard symmetric joint measurement will succeed. The major technical contribution of this paper is showing that this measurement also serves as a sample-efficient predictor for the principal eigenstate when used in the context of classical shadows.

Theorem 2 *Let $\rho = (1-\eta)\phi + \eta\sigma$ be an instance of principal eigenstate classical shadows. The standard joint measurement on n copies of state ρ succeeds with probability at least $(1-\eta)^{n-1}$. Conditioned on success of the measurement, there is an estimator $\hat{\phi}$ such that*

$$\mathbb{E}[\mathrm{Tr}(O\hat{\phi})] = \mathrm{Tr}(O\phi) + \frac{\eta \,\mathrm{Tr}(O(\sigma-\phi))}{n(1-\eta)} + \mathcal{O}\big(\eta^2/n\big)$$
$$\mathrm{Var}[\mathrm{Tr}(O\hat{\phi})] = \frac{\mathrm{Tr}(O^2)}{n^2} + \mathcal{O}\big(\eta^2 + 1/n\big)$$

for observables O with $\|O\|_\infty \leq 1$.

Notice that given Theorem 2, one can easily derive the sample complexity of the $\eta \leq 1/s^*$ regime given in Theorem 1 by invoking Chebyshev's inequality. Recall that the next two regimes are obtained by averaging, purification, or a combination thereof. We describe those procedures in Section 4. We give a detailed proof outline for Theorem 2 in Section 3, with full proofs in the appendix.

1.3. Related Work

The task of approximately preparing the principal eigenstate (a.k.a. purification) has a long history (Berthiaume et al., 1994; Barenco et al., 1997; Peres, 1999; Werner, 1998; Cirac et al., 1999; Lloyd et al., 2014; Fu, 2016; Childs et al., 2023). However, this task is costly achieving an η suppression that scales at most inversely in the number of copies of ρ (Werner, 1998; Cirac et al., 1999). This task is distinct from our work which aims to learn a classical description of the principal eigenstate to sufficient accuracy to permit future estimates of many expectation values.

More recently, focus has shifted from physical to virtual purification schemes. This more relevant body of previous work (Ekert et al., 2002; Cotler et al., 2019; Koczor, 2021b; Huggins et al., 2021; Czarnik et al., 2021), sometimes referred to as "virtual distillation", is one that directly learns an expectation value of a given observable with respect to $\rho_t := \rho^t / \mathrm{Tr}(\rho^t)$ for an unknown state ρ and an integer t. As t becomes large, ρ_t approaches the principal eigenstate of ρ connecting these techniques to our work. However, in contrast to our work, in these protocols, the observable is a part of the measurement circuit. So, for example, computing expected values for exponentially many observables would require exponential overhead in sample complexity, whereas our procedure in Theorem 1 requires only polynomial overhead. To be fair, there are certainly advantages to the virtual distillation setting for practical applications, most notably the fact that the sample complexity does not depend on properties of the observable such as its Frobenius norm.

Building on the virtual distillation program, Refs. Zhou and Liu (2022); Seif et al. (2023) consider learning non-linear functions of ρ such as $\mathrm{Tr}(O\rho^t)$; however, like our work, Refs. Zhou and Liu (2022); Seif et al. (2023) take a classical shadows style approach where measurements of the copies of ρ can be implemented without knowing the observables of

interest. Ref. Zhou and Liu (2022) shows that their sample complexities depend on B, the squared-Frobenius norm of the observable, achieving a sample complexity of $\mathcal{O}\left(\frac{(B+1)t}{\epsilon^2}\right)$ for the sub-procedure of estimating $\mathrm{Tr}(O\rho^t)$ to additive error ϵ. A straightforward calculation shows that solving the principal eigenstate classical shadows problem using these techniques for the estimator $\mathrm{Tr}(O\rho^t)/\mathrm{Tr}(\rho^t)$ results in much higher sample complexities for all regimes in which B is somewhat large (in particular, when $B > \eta/\sqrt{\epsilon}$). Consequently, our protocol is better for all regimes which do not rely on purification (i.e., $\eta \leq \sqrt{\epsilon}$). A more involved calculation shows that our protocol is still preferable in all but a handful of regimes, but they are harder to characterize (e.g., $B = 1$, $\epsilon = \eta^3$, and η sufficiently small). We leave a more thorough comparison of these techniques (including possible ways to combine them) to future work. We note that Ref. Zhou and Liu (2022) also considers the setting where O is a k-local observable. There, the sample complexity of estimating $\mathrm{Tr}(O\rho^t)$ is $\mathcal{O}\left(\frac{4^k t}{\epsilon^2}\right)$. Hence, for k-local observables where $B = 4^n \gg 4^k$, this protocol is preferable in several parameter regimes of interest.

Another related body of work surrounds classical shadows that are robust to noise in the measurement process itself (Chen et al., 2021; Koh and Grewal, 2022). In other words, those procedures work well when given a state that has been prepared with high fidelity, but are using low-fidelity measurements. Our procedure works well when given a low-fidelity state, but have measurements that can be performed with high fidelity.

1.4. Open Problems

Our work leaves open several new directions. Perhaps the most interesting is to explore variations of the principal eigenstate classical shadows problem. How do shadow estimation algorithms need to change to predict properties of the principal eigenstate, rather than the state itself? There are many possible variants worthy of consideration: when the class of observables is local (cf. Huang et al., 2020; Hakoshima et al., 2023); when the measurement procedure itself is faulty (cf. Chen et al., 2021; Koh and Grewal, 2022); when a low memory footprint is required (cf. Czarnik et al., 2021; Chen et al., 2022; Hakoshima et al., 2023); etc.

Another possible direction for future work is to generalize the principal eigenstate classical shadows problem to the top k eigenstates, rather than just the top eigenstate. When a complete description of the best rank-k approximator is needed, $O(kd/\epsilon^2)$ samples are sufficient by work of O'Donnell and Wright (2016) (where d is the dimension of the Hilbert space and ϵ is a bound on the trace distance to the optimal rank-k approximation), but once again, little is known in the classical shadows setting.

Finally, we ask whether or not our algorithm can be improved. When $\eta \leq 1/s^*$ (recall $s^* := \sqrt{B}/\epsilon + 1/\epsilon^2$), our algorithm obtains the same sample complexity as the $\eta = 0$ algorithm of Grier et al. (2022), which is provably optimal up to log factors. Therefore, our algorithm must also be optimal in that regime since we could always add noise in $\eta = 0$ setting if that improved the sample complexity. In the regime where $\eta > 1/s^*$, the optimality of our algorithm is unknown. However, one might suspect that there is the possibility for improvement since our algorithm does not measure all (or almost all) copies

of ρ at once, which is distinct from other optimal joint-measurement tomography algorithms (Haah et al., 2016; O'Donnell and Wright, 2016; Grier et al., 2022).

2. Abbreviated Preliminaries

We start with a condensed version of the preliminaries section in the appendix (cf. Appendix A). Let S_n denote the symmetric group of permutations on n elements.

Definition 3 (permutation operator) *Given a permutation $\pi \in \mathrm{S}_n$ (for $n \geq 1$), define a permutation operator $W_\pi \in \mathbb{C}^{d^n \times d^n}$ such that $W_\pi |x_1\rangle \cdots |x_n\rangle = \left|x_{\pi^{-1}(1)}\right\rangle \cdots \left|x_{\pi^{-1}(n)}\right\rangle$, and extend by linearity. That is, W_π acts on $(\mathbb{C}^d)^{\otimes n}$ by permuting the qudits, sending the qudit in position i to position $\pi(i)$.*

Definition 4 (symmetric subspace) *Define the* symmetric subspace *as the subspace of $(\mathbb{C}^d)^{\otimes n}$ fixed by the projector $\Pi_{\mathrm{sym}}^{(n)} = \frac{1}{n!}\sum_{\pi \in \mathrm{S}_n} W_\pi$, where $d_n = \binom{n+d-1}{d-1}$. Additionally, we have that $\Pi_{\mathrm{sym}}^{(n)} = d_n \int_\psi (|\psi\rangle\langle\psi|)^{\otimes n} \mathrm{d}\psi$ (e.g., Scott, 2006).*

Definition 5 *The* standard symmetric joint measurement *is a measurement on n qudits. It is defined by the POVM $\mathcal{M}_n = \{F_\psi\}_\psi \cup \{F_\perp\}$ with elements $F_\psi := d_n \cdot |\psi\rangle\langle\psi|^{\otimes n} \mathrm{d}\psi$, for all d-dimensional pure states ψ, proportional to the Haar measure $\mathrm{d}\psi$, plus a "failure" outcome $F_\perp := \mathbb{I} - \Pi_{\mathrm{sym}}^{(n)}$ for non-symmetric states.*

The measurement *fails* if we get outcome $F_\perp$, otherwise we say it *succeeds*. In what follows, we will let Ψ be the random variable representing the outcome of measuring $\rho^{\otimes n}$ with $\mathcal{M}_n$ where Ψ is 0 when the measurement fails and $|\psi\rangle\langle\psi|$ for measurement outcome ψ.

Note that standard techniques (employing t-designs, see Bajnok (1992); Hayashi et al. (2005); Bondarenko et al. (2013)) can be used to replace this measurement with a POVM with finitely many outcomes. Subsequently, the finite POVM can be compiled to a projective measurement Nielsen and Chuang (2010). See Ref. Grier et al. (2022) for more details on how to realize this measurement. All our sample complexity bounds hold under this replacement.

3. Outline of main theorem

We now give an outline of the proof of Theorem 2 to elucidate some of the key techniques. We refer the reader to the appendix for full proofs and details.

Let's begin with the most straightforward approach to proving Theorem 2—simply give exact expressions for the first and second moments of the standard symmetric measurement on $\rho^{\otimes n}$ conditioned on a successful[2] outcome:

Theorem 6 (Theorem 22 in Appendix B.1)

$$\mathbb{E}[\Psi \mid \text{success}] = \frac{\mathbb{I} + nM_1}{d+n}, \tag{1}$$

$$\mathbb{E}[\Psi^{\otimes 2} \mid \text{success}] = \frac{2\Pi_{\mathrm{sym}}^{(2)}}{(d+n)(d+n+1)}\left((\mathbb{I} + nM_1)^{\otimes 2} + \binom{n}{2} M_2 - n^2 M_1^{\otimes 2}\right), \tag{2}$$

2. As mentioned before, the fact that our measurement can fail (and output 0) is the consequence of our state not necessarily being in the symmetric subspace.

for mixed states $M_1 \propto \mathrm{Tr}_{1,\ldots,n-1}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})$ *and* $M_2 \propto \mathrm{Tr}_{1,\ldots,n-2}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})$.

Here, we are already forced to deviate from previous treatments (Huang et al., 2020; Grier et al., 2022). Notice that the expected value is not related by scaling and shifting by the identity to the unknown state ρ. Instead, the measurement's expectation is related to M_1, the partial trace of the projection of $\rho^{\otimes n}$ onto the symmetric subspace. As in (Grier et al., 2022), the proof of Theorem 6 relies on representation theory and properties of the symmetric subspace, but is considerably more streamlined by the use of Chiribella's theorem.

Ultimately, we will claim that M_1 is close to the principal eigenstate ϕ of ρ. So, the estimator $\hat{\phi}$ in Theorem 2 will be $((d+n)\Psi - \mathbb{I})/n$ conditioned on successful measurement. Two key challenges remain: first, we must show that M_1 is actually close to the principal eigenstate; second, we must bound the variance of our estimator. Unfortunately, the closed-form expressions for M_1 and M_2 are quite unwieldy.

To tackle these challenges, we reinterpret $\rho^{\otimes n}$ as a statistical mixture of states which are easier to analyze individually. To describe this decomposition, first let us write the unknown state as $\rho = \sum_{i=1}^{d} \lambda_i \Phi_i$ where $\lambda_1 \geq \cdots \geq \lambda_d$ and $\Phi_i := |\phi_i\rangle\langle\phi_i|$ are projectors onto the eigenstates. In the expansion of $\rho^{\otimes n}$, we will use vectors $\mathbf{e} = (e_1, \ldots, e_d) \in \mathbb{N}^d$ with $e_1 + \cdots + e_d = n$ to give counts for the different eigenstates of ρ. Now, we can define the mixed state

$$\sigma(\mathbf{e}) := \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} W_\pi \left(\Phi_1^{\otimes e_1} \otimes \cdots \otimes \Phi_d^{\otimes e_d}\right) W_\pi^\dagger$$

to be a symmetrized[3] version of the eigenstate $\Phi_1^{\otimes e_1} \otimes \cdots \otimes \Phi_d^{\otimes e_d}$. Using the short-hand expressions: $\mathbf{e}! := e_1! \cdots e_d!$, $\binom{n}{\mathbf{e}} := \frac{n!}{\mathbf{e}!}$, and $\underline{\lambda}^{\mathbf{e}} := \lambda_1^{e_1} \cdots \lambda_d^{e_d}$, we obtain our desired nice expansion of $\rho^{\otimes n}$:

Proposition 7 $\rho^{\otimes n} = \sum_{\mathbf{e}} \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \sigma(\mathbf{e})$.

We now interpret $\rho^{\otimes n}$ as statistical mixture of $\sigma(\mathbf{e})$ states where $\mathbf{e}$ is selected at random from the distribution with $\Pr[\mathbf{e}] = \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}}$, which we recognize as the multinomial distribution with n trials for d events with probabilities $\lambda_1, \ldots, \lambda_d$. However, the pertinent distribution for our calculations, which we name $\mathscr{D}$, is this multinomial *conditioned* on successful measurement. We arrive at new expressions (c.f. Theorem 28) for the first and second moments in Theorem 6 by expanding $\rho^{\otimes n}$ with this interpretation. For example, the first moment becomes

$$\mathbb{E}[\Psi \mid \text{success}] = \frac{\mathbb{I} + n\mathbb{E}_{\mathbf{e}\sim\mathscr{D}} M_1(\mathbf{e})}{d+n}$$

where mixed state $M_1(\mathbf{e}) \propto \mathrm{Tr}_{1,\ldots,n-1}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))$. This expansion has the potential to greatly simplify the calculation since $M_1(\mathbf{e})$ and $M_2(\mathbf{e})$ turn out to have surprisingly clean forms (c.f. theorems 30 and 31, respectively). Unfortunately, the distribution $\mathscr{D}$ is still quite complicated.

To circumvent this issue, our key observation is that the true distribution $\mathscr{D}$ is close to a distribution $\mathscr{D}'$ of independent geometric random variables. Technically, in $\mathscr{D}'$, each e_i with $i \geq 2$ is chosen independently from the geometric distribution with mean $\lambda_i/(\lambda_1 - \lambda_i)$ and e_1 is set to $n - (e_2 + \ldots + e_d)$.

3. To be clear, $\sigma(\mathbf{e})$ is typically *not* in the symmetric subspace since it is in general a mixed state.

Theorem 8 (Theorem 35 in Appendix B.5) $\|\mathscr{D} - \mathscr{D}'\|_{TV} \leq \left(\frac{1-\lambda_1}{\lambda_1}\right)^{n+1} \frac{\lambda_1}{2\lambda_1 - 1}$.

In other words, we can substitute $\mathscr{D}$ for $\mathscr{D}'$ without significant loss.[4] This geometric approximation dramatically simplifies many calculations, but nevertheless requires care to show it does not significantly affect the variance of our estimator, involving a sort of hybrid calculation where sometimes we assume the approximation and sometimes we do not. We leave these details to Appendix B.6 and Appendix B.7 for the first and second moments, respectively. Combining these pieces together completes the proof.

4. The compound estimation procedure

The full estimation procedure (to prove Theorem 1) uses our *measurement* in Theorem 2 as a black box, which is combined with *purification* of ρ before measurement, and *averaging* estimates from multiple measurements. We also require a step to estimate η from samples, to decide the η-regime of Theorem 1 and balance the *purification*, *measurement*, and *averaging* subroutines accordingly. Due to randomness in these subroutines, we will bound the expected number of samples.

In the purification step, we assume the existence of a black box which takes copies of ρ and creates a state ρ' as output. The number of copies consumed in this sub-procedure is a random variable with mean k but the output state is deterministic in the sense that for identical inputs ρ, identical outputs ρ' will be produced independent of the number of copies consumed. The purification procedure reduces the principal deviation, i.e., the deviation of ρ' satisfies:

$$\eta' = \mathcal{O}(\eta/k), \tag{3}$$

where $\eta < 1/2$ is the deviation of ρ. This result was shown to hold in the special case of $d = 2$ (Werner, 1998; Cirac et al., 1999). This result was later shown to hold in the general d setting in the special case where ρ is a convex combination of a pure quantum state and the maximally mixed state (Fu, 2016; Childs et al., 2023). Based on unpublished work, we claim that this result holds in greater generality: it applies to arbitrary mixed states in arbitrary dimension subject to $\eta < 1/2$. In our estimation procedure, we employ this result in the $\eta \in (\sqrt{\epsilon}, 1/2)$ regime.

In the measurement step, n copies of ρ' are consumed and an estimator $\hat{\phi}$ is output. This computation involves two steps. First, n copies of ρ' are measured using the standard symmetric joint measurement (cf. Definition 5) producing either a fail outcome or a classical description of a pure state Ψ. If a fail outcome is observed, the execution of the measurement sub-procedure fails on this instance resulting in n "wasted" copies of ρ'. We will be interested in the regime where the measurement sub-procedure will be executed many times with each having a constant probability of success hence, failures will at most contribute a constant factor to sample complexity. If the measurement succeeds, the measurement outcome is a classical description of a pure quantum state Ψ. An affine map is applied to produce $\hat{\phi}$, an estimator for M_1 and Φ_1:

$$\hat{\phi} = \frac{(d+n)\Psi - I}{n}. \tag{4}$$

4. This idea is similar, but not identical to a technique called "poissonization" (DasGupta, 2011).

This process is probabilistic so each call produces a different $\hat{\phi}$ with mean M_1 and variance given by Corollary 24.

In the averaging step, b independent estimates of M_1 are averaged to produce one improved estimate $\hat{\phi}(b)$. This has mean M_1 and a variance $\frac{1}{b}$ times that of $\hat{\phi}$. The estimator $\hat{\phi}(b)$ is an unbiased estimator of M_1 and a biased estimator of the principal eigenstate Φ_1. Using Theorem 2, for observables satisfying $\|O\|_\infty \leq 1$ the bias can be bounded by:

$$\beta = \left|\mathbb{E}[\mathrm{Tr}(O(\hat{\phi} - \phi))]\right| = \mathcal{O}\left(\frac{\eta'}{n}\right). \tag{5}$$

By ensuring that our estimator has bias $\mathcal{O}(\epsilon)$ and variance $\mathcal{O}(\epsilon^2)$, we employ Chebyshev's inequality to prove Theorem 1.

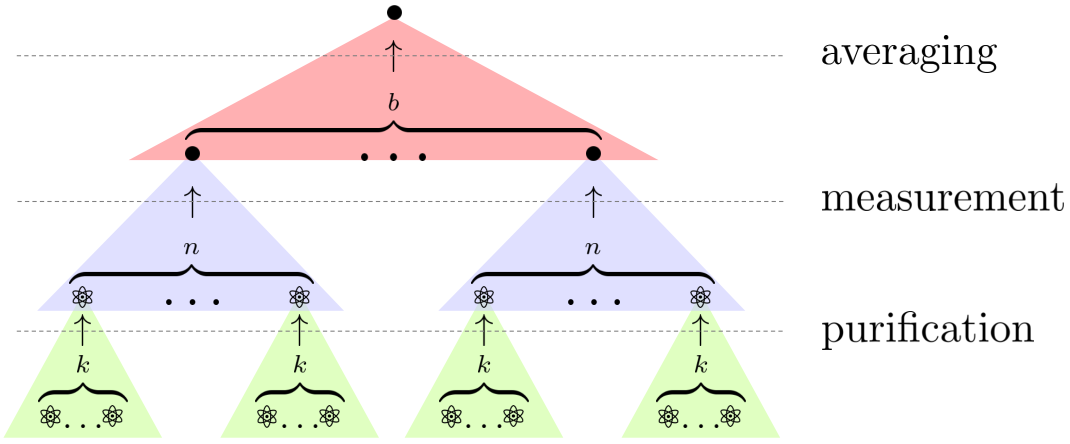


Figure 1: Our three step estimation procedure depicting the purification, measurement and averaging sub-procedures (from bottom to top). The purification procedure maps k quantum states to one quantum state (depicted by atom logos). The measurement procedure maps n quantum states to a classical description of an operator (depicted by the "•" symbol). The averaging procedure maps b classical descriptions to one classical description of an operator.

Figure 1 shows how these sub-procedures are combined to form our estimator $\hat{\phi}(b)$. For a given observable O, an expectation value can be estimated using $\mathrm{Tr}(O\hat{\phi}(b))$. With constant probability of failure (over the randomness of the measurement procedure), this produces an estimate of $\mathrm{Tr}(O\phi)$ up to additive error ϵ. By repeating this procedure and taking the median value over all repetitions, the probability of failure can be exponentially suppressed in the number of repetitions (Lugosi and Mendelson, 2019; Lerasle, 2019; Huang et al., 2020). We omit this standard "median-of-means" sub-procedure from our analysis, but note that it ensures that $\mathcal{O}(\log t)$ repetitions suffice to estimate the expectation value of t observables, $O_1, \ldots, O_t$, all to within additive error ϵ with high probability.

The remainder of this section will discuss the choice of parameters k, n and b and how the performance of our procedure compares to alternative approaches.

4.1. Choice of parameters

Three parameters (k, n, and b) define the algorithm, and control both the accuracy of our estimate and the expected number of samples of ρ used. We select values for these parameters based on the given values of B and ϵ, as well as η. Note that η is *not* given, but let us suppose we know it for now and come back to the problem of estimating η from samples after the theorem.

Theorem 9 *Given B, ϵ and η, the expected number of samples is minimized for the choice of k, n and b given in Table 1.*

The expectation is over the randomness in the purification and measurement procedures. The proof of Theorem 9 is given in Appendix E.

η	$\mathcal{O}(1/s^*)$	$\Omega(1/s^*) \cap \mathcal{O}(\sqrt{\epsilon})$	$\Omega(\sqrt{\epsilon})$
k	1	1	$\mathcal{O}\left(\frac{\eta}{\sqrt{\epsilon}}\right)$
n	$\mathcal{O}(s^*)$	$\mathcal{O}\left(\frac{1}{\eta}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$
b	1	$\mathcal{O}\left(\frac{B\eta^2+\eta}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{B}{\epsilon}+\frac{1}{\epsilon^{3/2}}\right)$
s	$\mathcal{O}(s^*)$	$\mathcal{O}\left(\frac{B\eta+1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{B\eta}{\epsilon^2}+\frac{\eta}{\epsilon^{5/2}}\right)$

Table 1: Choice of parameters k, n, b for the three regimes of η. Recall $s^* := \frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}$.

Last, we need a way to estimate η, since the choice of k, n, b are functions of η, either explicitly, or because they depend on the regime which is determined by η. Observe that a multiplicative approximation for η suffices since (i) in all three regimes, the complexity is linear in η hence any fixed multiplicative factor applied to η can be absorbed into the big-$\mathcal{O}$ constants, and (ii) adjacent regimes have the same complexity (up to constant factors) near the threshold, i.e., there is no "discontinuity" in the sample complexity with respect to η. Hence incorrect categorization of η-regime due to a multiplicative error still assigns a sample complexity that is equivalent to the sample complexity associated with the correct η-regime up to big-$\mathcal{O}$ constants. Finally, once we establish $\eta = \mathcal{O}(1/s^*)$ is in the first regime, we need no further estimate of η since k, n, b are functions of B and ϵ, not η.

Our information about η comes from joint measurements, specifically from when they *fail*. In Appendix B.4, Theorem 33, we show that the success probability of an n-sample measurement is bounded between $(1-\eta)^{n-1}$ and $(1-\eta)^{n-1}(1+\mathcal{O}(\eta^2))$. However, for a 2-sample measurement, we can be more specific:

$$\eta \leq \eta + \eta^2/2 \leq 1 - \Pr[\text{success}] \leq \eta + \eta^2 \leq 2\eta. \tag{6}$$

In other words, 2-sample measurements fail with some probability $p = \Theta(\eta)$ which is a multiplicative approximation for η. Hence, we can reduce to the problem of using many independent Bernoulli trials to estimate their failure probability. Indeed, we give an algorithm that does exactly this.

Theorem 10 *Let $r \geq 1$ be an integer. There is an algorithm which estimates the failure probability p of a Bernoulli trial, such that the algorithm (i) outputs a constant-factor multiplicative approximation of p, and (ii) makes $\mathcal{O}(r/p)$ samples of the Bernoulli trial, except with an $\exp(-\Theta(r))$ probability of failure.*

See Appendix F for the proof. We can get an arbitrarily low failure probability δ by taking $r = \mathcal{O}(\log(1/\delta))$, though Theorem 1 is stated for $\delta = \Omega(1)$ and hence $r = \mathcal{O}(1)$.

Finally, in the first regime of Theorem 1, $\eta = \mathcal{O}(1/s^*)$ can be small (or even zero!) and the algorithm in Theorem 10 would use too many samples if run to completion. Instead, we cut it off at $\Omega(r/p) = \Omega(r/s^*)$ samples, confident (w.p. $1 - \exp(-\Theta(r))$) that $\eta = \mathcal{O}(1/s^*)$, and then (conveniently) the parameters $k = b = 1$ and $n = \mathcal{O}(\sqrt{B}/\epsilon + 1/\epsilon^2)$ do not require an estimate of η.

4.2. Comparison to alternative approaches

There are two natural strategies to compare against. First, the original classical shadows paper uses (Huang et al., 2020) uses single-copy measurements which coincide with our single-copy measurement $\mathcal{M}_1$ (c.f. Definition 5). Below we give the optimal sample complexity within our *knb* framework when constrained to single-copy ($n = 1$) measurements.

Theorem 11 *With single-copy measurements and purification, we get sample complexity*

$$s = \begin{cases} \mathcal{O}(\frac{B}{\epsilon^2}) & \text{if } \eta \leq \epsilon, \\ \mathcal{O}(\frac{B\eta}{\epsilon^3}) & \text{if } \eta \geq \epsilon. \end{cases}$$

Second, we can turn to purification *before* repetition and averaging, i.e., set $b = 1$ in our framework. Again, the result is somewhat worse.

Theorem 12 *There is an algorithm which purifies and makes joint measurements (no averaging), having sample complexity*

$$s = \begin{cases} s^* & \text{if } \eta \leq 1/s^*, \\ \eta(s^*)^2 & \text{if } \eta \geq 1/s^*, \end{cases} \qquad \text{where } s^* = \frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}.$$

Since these fall within our *knb* framework, they cannot be any better than Theorem 1. Theorem 11 matches the performance of (Huang et al., 2020) initially, thus performing quadratically worse than Theorem 1 in the $B = \omega(1)$ regime. Theorem 11's performance then degrades by a factor of η/ϵ and compares poorly to the $\frac{B\eta+1}{\epsilon^2}$ performance in the middle regime of Theorem 1. Theorem 12 matches our performance for very small η, as expected, but then picks up a $\mathcal{O}\big(1/\epsilon^4\big)$ term which compares unfavorably with either the $\mathcal{O}\big(1/\epsilon^2\big)$ or $\mathcal{O}\big(1/\epsilon^{5/2}\big)$ terms in Theorem 1.

Acknowledgments

We thank Richard Kueng and anonymous referees for useful comments.

References

Bela Bajnok. Construction of spherical t-designs. *Geometriae Dedicata*, 43:167–179, 1992.

Adriano Barenco, André Berthiaume, David Deutsch, Artur Ekert, Richard Jozsa, and Chiara Macchiavello. Stabilization of quantum computations by symmetrization. *SIAM Journal on Computing*, 26(5):1541–1557, 1997. doi: 10.1137/S0097539796302452. URL https://doi.org/10.1137/S0097539796302452.

A. Berthiaume, D. Deutsch, and R. Jozsa. The stabilisation of quantum computations. In *Proceedings Workshop on Physics and Computation. PhysComp '94*, pages 60–62, 1994. doi: 10.1109/PHYCMP.1994.363698. URL https://doi.org/10.1109/PHYCMP.1994.363698.

Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1): 015004, 2022. doi: 10.1103/RevModPhys.94.015004. URL https://doi.org/10.1103/RevModPhys.94.015004.

Andriy Bondarenko, Danylo Radchenko, and Maryna Viazovska. Optimal asymptotic bounds for spherical designs. *Annals of Mathematics*, 178(2):443–452, 2013. ISSN 0003-486X. doi: 10.4007/annals.2013.178.2.2. URL http://dx.doi.org/10.4007/ANNALS.2013.178.2.2.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021. doi: 10.1038/s42254-021-00348-9. URL https://doi.org/10.1038/s42254-021-00348-9.

Senrui Chen, Wenjun Yu, Pei Zeng, and Steven T. Flammia. Robust shadow estimation. *PRX Quantum*, 2:030348, Sep 2021. doi: 10.1103/PRXQuantum.2.030348. URL https://link.aps.org/doi/10.1103/PRXQuantum.2.030348.

Sitan Chen, Jordan Cotler, Hsin-Yuan Huang, and Jerry Li. Exponential separations between learning with and without quantum memory. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 574–585. IEEE, 2022. doi: 10.1109/FOCS52979.2021.00063. URL https://doi.org/10.1109/FOCS52979.2021.00063.

Andrew M. Childs, Honghao Fu, Debbie Leung, Zhi Li, Maris Ozols, and Vedang Vyas. Streaming quantum state purification. *arXiv preprint arXiv:2309.16387*, 2023. doi: 10.48550/arXiv.2309.16387. URL https://doi.org/10.48550/arXiv.2309.16387.

Giulio Chiribella. On quantum estimation, quantum cloning and finite quantum de Finetti theorems. In *Theory of Quantum Computation, Communication, and Cryptography: 5th Conference, TQC 2010, Leeds, UK, April 13-15, 2010, Revised Selected Papers 5*, pages 9–25. Springer, 2011. doi: 10.1007/978-3-642-18073-6_2. URL https://doi.org/10.1007/978-3-642-18073-6_2.

J Ignacio Cirac, A. K. Ekert, and Chiara Macchiavello. Optimal purification of single qubits. *Physical review letters*, 82(21):4344, 1999. doi: 10.1103/PhysRevLett.82.4344. URL https://doi.org/10.1103/PhysRevLett.82.4344.

Jordan Cotler, Soonwon Choi, Alexander Lukin, Hrant Gharibyan, Tarun Grover, M. Eric Tai, Matthew Rispoli, Robert Schittko, Philipp M. Preiss, Adam M. Kaufman, Markus Greiner, Hannes Pichler, and Patrick Hayden. Quantum virtual cooling. *Phys. Rev. X*, 9: 031013, Jul 2019. doi: 10.1103/PhysRevX.9.031013. URL https://link.aps.org/doi/10.1103/PhysRevX.9.031013.

Piotr Czarnik, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. Qubit-efficient exponential suppression of errors. *arXiv preprint arXiv:2102.06056*, 2021. doi: 10.48550/arXiv.2102.06056. URL https://doi.org/10.48550/arXiv.2102.06056.

Anirban DasGupta. *Probability for statistics and machine learning: fundamentals and advanced topics*. Springer, 2011. doi: 10.1007/978-1-4419-9634-3. URL https://doi.org/10.1007/978-1-4419-9634-3.

Artur K. Ekert, Carolina Moura Alves, Daniel K. L. Oi, Michał Horodecki, Paweł Horodecki, and L. C. Kwek. Direct estimations of linear and nonlinear functionals of a quantum state. *Phys. Rev. Lett.*, 88:217901, May 2002. doi: 10.1103/PhysRevLett.88.217901. URL https://link.aps.org/doi/10.1103/PhysRevLett.88.217901.

Honghao Fu. Quantum state purification. Thesis: University of Waterloo, 2016. URL http://hdl.handle.net/10012/10706.

Daniel Grier, Hakop Pashayan, and Luke Schaeffer. Sample-optimal classical shadows for pure states. *arXiv preprint arXiv:2211.11810*, 2022. doi: 10.48550/arXiv.2211.11810. URL https://doi.org/10.48550/arXiv.2211.11810.

Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 913–925, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325. doi: 10.1145/2897518.2897585. URL https://doi.org/10.1145/2897518.2897585.

Hideaki Hakoshima, Suguru Endo, Kaoru Yamamoto, Yuichiro Matsuzaki, and Nobuyuki Yoshioka. Localized virtual purification. *arXiv preprint arXiv:2308.13500*, 2023. doi: 10.48550/arXiv.2308.13500. URL https://doi.org/10.48550/arXiv.2308.13500.

Aram W Harrow. The church of the symmetric subspace. *arXiv preprint arXiv:1308.6595*, 2013. doi: 10.48550/arXiv.1308.6595. URL https://doi.org/10.48550/arXiv.1308.6595.

A Hayashi, T Hashimoto, and M Horibe. Reexamination of optimal quantum state estimation of pure states. *Physical review A*, 72(3):032325, 2005.

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

2020. ISSN 1745-2481. doi: 10.1038/s41567-020-0932-7. URL https://doi.org/10.1038/s41567-020-0932-7.

William J Huggins, Sam McArdle, Thomas E O'Brien, Joonho Lee, Nicholas C Rubin, Sergio Boixo, K Birgitta Whaley, Ryan Babbush, and Jarrod R McClean. Virtual distillation for quantum error mitigation. *Physical Review X*, 11(4):041036, 2021. doi: 10.1103/PhysRevX.11.041036. URL https://link.aps.org/doi/10.1103/PhysRevX.11.041036.

Michael Keyl and Reinhard F Werner. The rate of optimal purification procedures. In *Annales Henri Poincare*, volume 2, pages 1–26. Springer, 2001. doi: 10.1007/PL00001027. URL https://doi.org/10.1007/PL00001027.

Bálint Koczor. The dominant eigenvector of a noisy quantum state. *New Journal of Physics*, 23(12):123047, 2021a. doi: 10.1088/1367-2630/ac37ae. URL https://dx.doi.org/10.1088/1367-2630/ac37ae.

Bálint Koczor. Exponential error suppression for near-term quantum devices. *Physical Review X*, 11(3):031057, 2021b. doi: 10.1103/PhysRevX.11.031057. URL https://link.aps.org/doi/10.1103/PhysRevX.11.031057.

Dax Enshan Koh and Sabee Grewal. Classical shadows with noise. *Quantum*, 6: 776, 2022. doi: 10.22331/q-2022-08-16-776. URL https://doi.org/10.22331/q-2022-08-16-776.

M. Lerasle. Lecture notes: Selected topics on robust statistical learning theory. *arXiv preprint arXiv:1908.10761*, 2019. doi: 10.48550/arXiv.1908.10761. URL https://doi.org/10.48550/arXiv.1908.10761.

Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nature Physics*, 10(9):631–633, Jul 2014. ISSN 1745-2481. doi: 10.1038/nphys3029. URL http://dx.doi.org/10.1038/nphys3029.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comp. Math.*, 19:1145–1190, 2019. doi: 10.1007/s10208-019-09427-x. URL https://doi.org/10.1007/s10208-019-09427-x.

Serge Massar and Sandu Popescu. Optimal extraction of information from finite quantum ensembles. *Physical review letters*, 74(8):1259, 1995. doi: 10.1103/PhysRevLett.74.1259. URL https://link.aps.org/doi/10.1103/PhysRevLett.74.1259.

Antonio Anna Mele. Introduction to Haar measure tools in quantum information: A beginner's tutorial. *arXiv preprint arXiv:2307.08956*, 2023. doi: 10.48550/arXiv.2307.08956. URL https://doi.org/10.48550/arXiv.2307.08956.

Michael A Nielsen and Isaac L Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010. doi: 10.1017/CBO9780511976667. URL https://doi.org/10.1017/CBO9780511976667.

Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 899–912, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325. doi: 10.1145/2897518.2897544. URL https://doi.org/10.1145/2897518.2897544.

Asher Peres. Error symmetrization in quantum computers. *International Journal of Theoretical Physics*, 38(3):799–805, Mar 1999. ISSN 1572-9575. doi: 10.1023/A:1026648717079. URL https://doi.org/10.1023/A:1026648717079.

Daniel A Roberts and Beni Yoshida. Chaos and complexity by design. *Journal of High Energy Physics*, 2017(4):1–64, 2017. doi: 10.1007/JHEP04(2017)121. URL https://doi.org/10.1007/JHEP04(2017)121.

Andrew J Scott. Tight informationally complete quantum measurements. *Journal of Physics A: Mathematical and General*, 39(43):13507, 2006. doi: 10.1088/0305-4470/39/43/009. URL https://dx.doi.org/10.1088/0305-4470/39/43/009.

Alireza Seif, Ze-Pei Cian, Sisi Zhou, Senrui Chen, and Liang Jiang. Shadow distillation: Quantum error mitigation with classical shadows for near-term quantum processors. *PRX Quantum*, 4:010303, Jan 2023. doi: 10.1103/PRXQuantum.4.010303. URL https://link.aps.org/doi/10.1103/PRXQuantum.4.010303.

Reinhard F Werner. Optimal cloning of pure states. *Physical Review A*, 58(3):1827, 1998. doi: 10.1103/PhysRevA.58.1827. URL https://link.aps.org/doi/10.1103/PhysRevA.58.1827.

John Wright. Lecture 22: Introduction to quantum tomography, 2015. URL https://www.cs.cmu.edu/~odonnell/quantum15/lecture22.pdf.

Kaoru Yamamoto, Suguru Endo, Hideaki Hakoshima, Yuichiro Matsuzaki, and Yuuki Tokunaga. Error-mitigated quantum metrology via virtual purification. *Phys. Rev. Lett.*, 129:250503, 2022. doi: 10.1103/PhysRevLett.129.250503. URL https://link.aps.org/doi/10.1103/PhysRevLett.129.250503.

You Zhou and Zhenhuan Liu. A hybrid framework for estimating nonlinear functions of quantum states. *arXiv preprint arXiv:2208.08416*, 2022. doi: 10.48550/arXiv.2208.08416. URL https://doi.org/10.48550/arXiv.2208.08416.

Appendix A. Preliminaries

Let $\mathbb{C}$ be the set of complex numbers and $\mathbb{C}^d$ the space of d-dimensional complex vectors.

Definition 13 (Operator spaces) *Given a Hilbert space V, let $\mathcal{L}(V)$ denote the set of linear operators from V to V. Let*

$$\mathrm{Dens}(V) = \{A \in \mathcal{L}(V) : \mathrm{Tr}(A) = 1, A \succeq 0\}$$

be the set of density matrices *which have trace 1 and are positive semidefinite.*

A.1. Symmetric and exchangeable operators

Let S_n denote the symmetric group of permutations on n elements.

Definition 14 (permutation operator) *Given a permutation $\pi \in \mathrm{S}_n$ (for $n \geq 1$), define a permutation operator $W_\pi \in \mathbb{C}^{d^n \times d^n}$ such that*

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

and extend by linearity. That is, W_π acts on $(\mathbb{C}^d)^{\otimes n}$ by permuting the qudits, sending the qudit in position i to position $\pi(i)$.

Definition 15 (symmetric and exchangeable) *Let $\rho \in \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big)$. We say ρ is* exchangeable *if $\rho = W_\pi \rho W_\pi^{-1}$ for all $\pi \in \mathrm{S}_n$, and ρ is* symmetric *if $\rho = W_\pi \rho W_\sigma$ for all $\pi, \sigma \in \mathrm{S}_n$.*

Put another way, exchangeable operators *commute* with W_π or $\Pi_{\mathrm{sym}}^{(n)}$, whereas symmetric operators *absorb* W_π and $\Pi_{\mathrm{sym}}^{(n)}$. Naturally, an *exchangeable state* or *symmetric state* is a state (i.e., a density matrix) which is also exchangeable or symmetric respectively.

Definition 16 (symmetric subspace) *The* symmetric subspace *of a system of n qudits of dimension d is the set of all symmetric operators.*[5] *Let d_n to denote the dimension of the symmetric subspace and define $\Pi_{\mathrm{sym}}^{(n)}$ to be the projector onto it (notationally omitting the dependence on d, the dimension of the qudit, which is typically fixed for our purposes).*

It's worth noting that states in the symmetric subspace are also exchangeable, but exchangeable are not necessarily in the symmetric subspace—take, for example, the maximally mixed state. We have two characterizations of the symmetric subspace.

Fact 17 *For all $n \geq 0$, $\Pi_{\mathrm{sym}}^{(n)} = \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} W_\pi$, and $d_n = \binom{n+d-1}{d-1}$.*

The integral of $|\psi\rangle\langle\psi|$ over the Haar measure is known from, e.g., Scott (2006).

Lemma 18

$$d_n \int_\psi (|\psi\rangle\langle\psi|)^{\otimes n} \mathrm{d}\psi = \Pi_{\mathrm{sym}}^{(n)} = \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} W_\pi$$

where $\Pi_{\mathrm{sym}}^{(n)}$ is the projector onto the symmetric subspace and W_π is the operator that permutes n qudits by an n-element permutation $\pi \in \mathrm{S}_n$.

5. One can check that symmetric operators are closed under addition and scalar multiplication, and thus this is subspace.

Recall that for a system $V = V_A \otimes V_B$, the *partial trace* $\mathrm{Tr}_A \colon \mathcal{L}(V_A \otimes V_B) \to \mathcal{L}(V_B)$ is the unique superoperator such that

$$\operatorname*{Tr}_A(\rho_A \otimes \rho_B) = \mathrm{Tr}(\rho_A)\rho_B$$

for all $\rho_A \in \mathcal{L}(V_A)$ and $\rho_B \in \mathcal{L}(V_B)$. We use the notation $\mathrm{Tr}_{n\to k}$ to represent the map from $\mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big)$ to $\mathcal{L}\big((\mathbb{C}^d)^{\otimes k}\big)$ given by

$$\operatorname*{Tr}_{n\to k}(A) := \operatorname*{Tr}_{1,\ldots,n-k}(\Pi^{(n)}_{\mathrm{sym}} A) \tag{7}$$

That is, $\mathrm{Tr}_{n\to k}$ reduces an n-qudit state to k qudits symmetrically.

Appendix B. The Standard Symmetric Joint Measurement on almost-pure states

Let us recall the measurement used in pure state classical shadows Grier et al. (2022) and many other learning tasks.

Definition 19 *The* standard symmetric joint measurement *is a measurement on n qudits. It is defined by the POVM $\mathcal{M}_n = \{F_\psi\}_\psi \cup \{F_\perp\}$ with elements*

$$F_\psi := d_n \cdot |\psi\rangle\langle\psi|^{\otimes n} \mathrm{d}\psi,$$

for all d-dimensional pure states ψ, proportional to the Haar measure $\mathrm{d}\psi$, plus a "failure" outcome $F_\perp := \mathbb{I} - \Pi^{(n)}_{\mathrm{sym}}$ for non-symmetric states.

The measurement *fails* if we get outcome $F_\perp$, otherwise we say it *succeeds*. In what follows, we will let Ψ be a random variable over $\mathcal{L}\big(\mathbb{C}^d\big)$ representing the outcome of measuring $\rho^{\otimes n}$ with $\mathcal{M}_n$ where

$$\Psi = \begin{cases} 0 & \text{for outcome } \perp, \\ |\psi\rangle\langle\psi| & \text{for outcome } \psi. \end{cases}$$

When the measurement succeeds, we construct an estimator $\hat{\phi} = \frac{(d+n)\Psi - \mathbb{I}}{n}$ from this random variable. We quantify the performance of $\hat{\phi}$ in the theorem below, but it is close enough to Φ_1 to be useful in solving the Principal Eigenstate Classical Shadows problem.

Theorem 2 *Let $\rho = (1-\eta)\phi + \eta\sigma$ be an instance of principal eigenstate classical shadows. The standard joint measurement on n copies of state ρ succeeds with probability at least $(1-\eta)^{n-1}$. Conditioned on success of the measurement, there is an estimator $\hat{\phi}$ such that*

$$\mathbb{E}[\mathrm{Tr}(O\hat{\phi})] = \mathrm{Tr}(O\phi) + \frac{\eta \,\mathrm{Tr}(O(\sigma - \phi))}{n(1-\eta)} + \mathcal{O}\big(\eta^2/n\big)$$
$$\mathrm{Var}[\mathrm{Tr}(O\hat{\phi})] = \frac{\mathrm{Tr}(O^2)}{n^2} + \mathcal{O}\big(\eta^2 + 1/n\big)$$

for observables O with $\|O\|_\infty \leq 1$.

The proof of this theorem is one of the main technical contributions of this paper, and we spend the rest of this section proving it. Before we begin, let us give a detailed outline of the structure of the proof, to help orient the reader.

First, the strategy of *measuring* copies of a state and *preparing* a fixed state for each outcome is known as a *measure-and-prepare channel*. Our particular case—measure with $\mathcal{M}_n$ and prepare Ψ—is especially well studied. In Appendix C, we adapt a result of Chiribella Chiribella (2011) (Theorem 48) about this channel to get formulas for $\mathbb{E}[\Psi]$, $\mathbb{E}[\Psi^{\otimes 2}]$, and $\mathrm{Var}[\Psi]$ (Theorem 22).

However, the formulas for $\mathbb{E}[\Psi]$, $\mathbb{E}[\Psi^{\otimes 2}]$, and $\mathrm{Var}[\Psi]$ are in terms two linear operators,

$$M_1 := \frac{\mathrm{Tr}_{1,\dots,n-1}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})}{\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})}, \qquad M_2 := \frac{\mathrm{Tr}_{1,\dots,n-2}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})}{\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})}.$$

While it is easy to write, e.g., $\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})$, it is not so easy to bound it in terms of the principal eigenvalue (λ_1) or the deviation (η). For instance, a natural approach is to expand $\Pi^{(n)}_{\mathrm{sym}}$ as a sum of permutations (by Fact 17), and calculate $\mathrm{Tr}(W_\pi\rho^{\otimes n}) = \prod_{C\in\pi}\mathrm{Tr}(\rho^{|C|})$ where the product is over cycles C of π. This argument establishes that $\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\rho^{\otimes n})$ is a symmetric polynomial in the spectrum $\lambda_1,\dots,\lambda_n$ of ρ, but the dependence on η is difficult to bound. We need a different approach.

Our solution is to expand $\rho^{\otimes n}$ in the basis $\{\Phi_{a_1}\otimes\cdots\otimes\Phi_{a_n} : a_1,\dots,a_n\in[d]\}$ where $\Phi_1,\dots,\Phi_n\in\mathrm{Dens}(\mathbb{C}^d)$ are the eigenvectors of ρ. Actually, since $\rho^{\otimes n}$ is symmetric, we are more concerned with the number of occurrences $e_i = \#\{j : a_j = i\}$ of each factor Φ_i, which constitute a *type* $\mathbf{e} = (e_1,\dots,e_d)$ of basis state. We define $\sigma(\mathbf{e})$ as the average of all states of type $\mathbf{e}$. Theorem 28 shows that M_1 and M_2 are expectations of $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ below, where $\mathbf{e}$ is sampled from some distribution $\mathscr{D}$.

$$M_1^{\mathbf{e}} := \frac{\mathrm{Tr}_{1,\dots,n-1}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))}{\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))}, \qquad M_2^{\mathbf{e}} := \frac{\mathrm{Tr}_{1,\dots,n-2}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))}{\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))}.$$

We have reduced the computation of M_1 and M_2 to the computation of $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ for a vast set of $\mathbf{e}$, but is this really progress? Yes (!), because the following trace is either 0 or 1, depending on the permutation π, and similar facts are true of the partial traces we need.

Fact 20

$$\mathrm{Tr}(W_\pi(\Phi_{a_1}\otimes\cdots\otimes\Phi_{a_n})) = \begin{cases} 0 & \text{if } a_{\pi(j)} \neq a_j \text{ for some } j\in[n], \\ 1 & \text{otherwise.}\end{cases}$$

This transforms the computation of $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ into a combinatorial problem: we can compute, e.g., $\mathrm{Tr}(\Pi^{(n)}_{\mathrm{sym}}\sigma(\mathbf{e}))$ by *counting* how many permutations contribute 1 (rather than 0). Using this approach, we compute a probability of success, $Z^{\mathbf{e}}$, in Theorem 29, and $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ in Theorems 30 and 31 respectively.

In principle, we now have concrete expressions for $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ in terms of $\mathbf{e}$ and $\{\Phi_i\}_{i=1}^d$, and it only remains to compute expectations over $\mathbf{e}\sim\mathscr{D}$. Here the distribution $\mathscr{D}$ puts

weight on $\mathbf{e}$ proportional to $\lambda_1^{e_1} \cdots \lambda^{e_d}$. Despite the simplicity of this probability mass function, we could find no closed form for $\mathbb{E}_{\mathbf{e}\in\mathscr{D}}[e_i]$ or $\mathbb{E}_{\mathbf{e}\in\mathscr{D}}[e_i e_j]$. This last technical hurdle forces us to approximate: we introduce a distribution $\mathscr{D}'$ for which we *can* compute $\mathbb{E}_{\mathbf{e}\in\mathscr{D}'}[e_i]$ and $\mathbb{E}_{\mathbf{e}\in\mathscr{D}'}[e_i e_j]$, and show that is it close to $\mathscr{D}$ when η is sufficiently small.

We finish the proof by combining the pieces. The expectation of $\hat{\rho}$, for example, is finally completed in Corollary 39. The corollary uses various theorems to bound the distance to the expectation under $\mathscr{D}'$, which is given by Theorem 38. Theorem 38 is derived from Corollary 24 to claim the expectation of the estimator (called $\hat{M}_1$ rather than $\hat{\rho}$ in that section) is M_1, Theorem 28 to expand M_1 as a distribution over $M_1^{\mathbf{e}}$, Theorem 30 to turn $M_1^{\mathbf{e}}$ into actual e_i and Φ_i terms, and Proposition 50 to evaluate those for the geometric random variables defining $\mathscr{D}'$.

B.1. Chiribella's Theorem: Moments from Partial Traces

The measurement we use $(\mathcal{M}_n)$ has applications to other pure state learning tasks, so there is already a result characterizing the moments of the outcome (i.e., $\mathbb{E}[\Psi^{\otimes k}]$), and in particular the mean and variance. Specifically, in Appendix C, we take a result of Chiribella Chiribella (2011) and repackage it into the following.

Theorem 21 *Fix integers $n, k \geq 0$, let $A \in \mathrm{Dens}\big((\mathbb{C}^d)^{\otimes n}\big)$ be an exchangeable n-qudit state, and let Ψ be the outcome of measuring A with $\mathcal{M}_n$.*

$$\mathbb{E}[\Psi^{\otimes k} \mid \text{success}] = \frac{1}{(d+n)_k} \Pi_{\text{sym}}^{(k)} \left(\sum_{s=0}^{k} \binom{n}{s} \binom{k}{s} \left(\frac{\mathrm{Tr}_{n\to s}(A)}{\mathrm{Tr}_{n\to 0}(A)} \otimes \mathbb{I}^{\otimes k-s} \right) \right) \Pi_{\text{sym}}^{(k)} \tag{8}$$

In other words, to compute $\mathbb{E}[\Psi^{\otimes k} \mid \text{success}]$ we only need $\mathrm{Tr}_{n\to s}(A) = \mathrm{Tr}_{[n-s]}(\Pi_{\text{sym}}^{(n)} A)$ for $0 \leq s \leq k$. More specifically, the important data about the state are $M_0(A), \ldots, M_k(A)$ where we define the function[6]

$$M_k(A) := \frac{\mathrm{Tr}_{n\to k}(A)}{\mathrm{Tr}_{n\to 0}(A)}.$$

We further abbreviate $M_k(A)$ to just M_k when the state is understood.

Since we aim to compute the mean and variance of Ψ, i.e., $\mathbb{E}[\Psi \mid \text{success}]$ and

$$\mathrm{Var}[\Psi \mid \text{success}] = \mathbb{E}[\Psi \mid \text{success}^{\otimes 2}] - \mathbb{E}[\Psi \mid \text{success}]^{\otimes 2},$$

we only apply Theorem 21 with $k = 1, 2$. Below, we specialize Theorem 21 to these two cases, using our new M_1, M_2 notation.

Theorem 22 *Let A be an exchangeable n-qudit state, and let Ψ be a random variable for the outcome of measuring A with $\mathcal{M}_n$.*

$$\mathbb{E}[\Psi \mid \text{success}] = \frac{\mathbb{I} + nM_1}{d+n}, \tag{9}$$

$$\mathbb{E}[\Psi^{\otimes 2} \mid \text{success}] = \frac{2\Pi_{\text{sym}}^{(2)}}{(d+n)(d+n+1)} \left((\mathbb{I} + nM_1)^{\otimes 2} + \binom{n}{2} M_2 - n^2 M_1^{\otimes 2} \right), \tag{10}$$

6. This is technically a partial function because of potential division by 0, but $\mathrm{Tr}_{n\to 0}(A) = \mathrm{Tr}(\Pi_{\text{sym}}^{(n)} A) = \Pr[\text{success}]$, so this is only a problem if the measurement always fails, and then we have bigger problems.

$$\mathrm{Var}[\Psi \mid \text{success}] = \frac{W_{(1\,2)}(\mathbb{I} + nM_1 \otimes \mathbb{I} + n\mathbb{I} \otimes M_1)}{(d+n)(d+n+1)} + \frac{n(n-1)M_2 - n^2 M_1^{\otimes 2}}{(d+n)(d+n+1)} - \frac{(\mathbb{I} + nM_1)^{\otimes 2}}{(d+n)^2(d+n+1)}, \tag{11}$$

where the expectation and variance are over the randomness in the measurement.

Proof $\mathbb{E}[\Psi \mid \text{success}]$ and $\mathbb{E}[\Psi^{\otimes 2} \mid \text{success}]$ are immediate from Theorem 21. For the variance, we start with the definition:

$$\begin{aligned}\mathrm{Var}[\Psi \mid \text{success}] &= \mathbb{E}[\Psi^{\otimes 2} \mid \text{success}] - \mathbb{E}[\Psi \mid \text{success}]^{\otimes 2} \\ &= \frac{2\Pi^{(2)}_{\mathrm{sym}}}{(d+n)(d+n+1)}\left((\mathbb{I} + nM_1)^{\otimes 2} + \binom{n}{2} M_2 - n^2 M_1^{\otimes 2}\right) - \left(\frac{\mathbb{I} + nM_1}{d+n}\right)^{\otimes 2}.\end{aligned}$$

Recall that $2\Pi^{(n)}_{\mathrm{sym}} = W_{(1)(2)} + W_{(1\,2)}$, so first and foremost there is a near-cancellation of two terms:

$$\frac{1}{(d+n)(d+n+1)} W_{(1)(2)}(\mathbb{I} + nM_1)^{\otimes 2} - \left(\frac{\mathbb{I} + nM_1}{d+n}\right)^{\otimes 2} = -\frac{(\mathbb{I} + nM_1)^{\otimes 2}}{(d+n)^2(d+n+1)}.$$

The remaining terms are

$$\frac{1}{(d+n)(d+n+1)}\left(\binom{n}{2} M_2 - n^2 M_1^{\otimes 2}\right) + \frac{W_{(1\,2)}}{(d+n)(d+n+1)}\left((\mathbb{I} + nM_1)^{\otimes 2} + \binom{n}{2} M_2 - n^2 M_1^{\otimes 2}\right).$$

Among the $W_{(1\,2)}$ terms, $(\mathbb{I} + nM_1)^{\otimes 2}$ cancels with $n^2 M_1^{\otimes 2}$ leaving

$$\frac{W_{(1\,2)}}{(d+n)(d+n+1)}(\mathbb{I} + nM_1 \otimes \mathbb{I} + n\mathbb{I} \otimes M_1).$$

Moreover, M_2 is invariant under $W_{(1\,2)}$ (i.e., $M_2 = W_{(1\,2)} M_2$), the two M_2 terms combine into $\frac{n(n-1)}{(d+n)(d+n+1)} M_2$, which we group with the left over $-n^2 M_1^{\otimes 2}$ term.

Altogether, the variance is

$$\mathrm{Var}[\Psi \mid \text{success}] = \frac{W_{(1\,2)}(\mathbb{I} + nM_1 \otimes \mathbb{I} + n\mathbb{I} \otimes M_1)}{(d+n)(d+n+1)} + \frac{n(n-1)M_2 - n^2 M_1^{\otimes 2}}{(d+n)(d+n+1)} - \frac{(\mathbb{I} + nM_1)^{\otimes 2}}{(d+n)^2(d+n+1)}.$$

■

B.2. Estimator $\hat{M}_1$

Consider the mean of Ψ,

$$\mathbb{E}[\Psi \mid \text{success}] = \frac{\mathbb{I} + nM_1}{d+n},$$

from Theorem 22. This is a convex combination of $\mathbb{I}/d$, the maximally mixed state, and M_1, which we observe is also a state.

Lemma 23 *For any exchangeable state $A \in \mathrm{Dens}\big((\mathbb{C}^d)^{\otimes n}\big)$ such that $\mathrm{Tr}_{n\to 0}(A) \neq 0$, we have $M_k(A) \in \mathrm{Dens}\big((\mathbb{C}^d)^{\otimes k}\big)$ for all $k \geq 0$.*

Proof If $M_k(A)$ is not positive semi-definite, then there is a state σ that witnesses its negativity (i.e., $\mathrm{Tr}(M_k(A)\sigma) < 0$). Then $\mathbb{I}\otimes\sigma$ witnesses the negativity of $\Pi_{\mathrm{sym}}^{(n)} A$ ($\mathrm{Tr}((\Pi_{\mathrm{sym}}^{(n)} A)(\mathbb{I}\otimes\sigma)) < 0$), a contradiction.

For the trace, observe that

$$\mathrm{Tr}(M_k(A)) = \frac{\mathrm{Tr}(\mathrm{Tr}_{[n-k]}(\Pi_{\mathrm{sym}}^{(n)} A))}{\mathrm{Tr}_{n\to 0}(A)} = \frac{\mathrm{Tr}(\Pi_{\mathrm{sym}}^{(n)} A)}{\mathrm{Tr}_{n\to 0}(A)} = \frac{\mathrm{Tr}_{n\to 0}(A)}{\mathrm{Tr}_{n\to 0}(A)} = 1.$$

■

In any case, clearly M_1 is the part of the estimator that we are interested in, the "signal" among the "noise". It is standard practice in classical shadows protocols to "invert" the channel to isolate the component of interest (in our case, M_1). That is, define an estimator[7] $\hat{M}_1 := \frac{1}{n}[(d+n)\Psi - \mathbb{I}]$ so that

$$\mathbb{E}[\hat{M}_1] = \tfrac{1}{n}[(d+n)\,\mathbb{E}[\Psi \mid \text{success}] - \mathbb{I}] = M_1.$$

It is important that *we only define $\hat{M}_1$ when the measurement succeeds*—we throw away any failed measurements.[8]

Corollary 24 (Estimator mean and variance in terms of M_1, M_2) *Suppose we measure a state $A \in \mathrm{Dens}\big((\mathbb{C}^d)^{\otimes n}\big)$, the measurement succeeds, and we produce estimator $\hat{M}_1$ as described. Let O be a Hermitian observable. The expectation and variance over the randomness of the measurement outcome,*

$$\mathop{\mathbb{E}}_{\text{meas}}[\hat{M}_1] = M_1 \tag{12}$$

$$\mathop{\mathrm{Var}}_{\text{meas}}[\mathrm{Tr}(O\hat{M}_1)] \le \frac{\mathrm{Tr}(O^2)}{n^2} + \frac{2\|O\|_\infty^2}{n} + \frac{n-1}{n}\,\mathrm{Tr}(O^{\otimes 2}M_2) - \mathrm{Tr}(OM_1)^2, \tag{13}$$

are functions of $M_1 := M_1(A)$, $M_2 := M_2(A)$.

Proof We have already seen $\mathbb{E}[\hat{M}_1]$, and

$$\mathrm{Var}[\hat{M}_1] = \mathrm{Var}\big[\tfrac{1}{n}[(d+n)\Psi - \mathbb{I}]\big] = \mathrm{Var}[\tfrac{d+n}{n}\Psi] = \tfrac{(d+n)^2}{n^2}\,\mathrm{Var}[\Psi].$$

We can substitute in (11) from Theorem 22 to get an expression for $\mathrm{Var}_{\text{meas}}[\hat{M}_1]$.

$$\mathop{\mathrm{Var}}_{\text{meas}}[\hat{M}_1] = \frac{d+n}{d+n+1}\left(\frac{W_{(1\,2)}(\mathbb{I}+nM_1\otimes\mathbb{I}+n\mathbb{I}\otimes M_1)}{n^2} + \frac{n-1}{n}M_2 - M_1^{\otimes 2} - \frac{(\mathbb{I}+nM_1)^{\otimes 2}}{n^2(d+n)^2}\right)$$

We remind the reader that $\mathrm{Var}[\hat{M}_1]$ is a 2-qudit linear operator, and to get the variance of $\mathrm{Tr}(O\hat{M}_1)$, we need to take the trace with $O^{\otimes 2}$ since

$$\begin{aligned}\mathrm{Var}[\mathrm{Tr}(O\hat{M}_1)] &= \mathbb{E}[\mathrm{Tr}(O\hat{M}_1)^2] - \mathbb{E}[\mathrm{Tr}(O\hat{M}_1)]^2 \\ &= \mathrm{Tr}(O^{\otimes 2}\,\mathbb{E}[\hat{M}_1^{\otimes 2}]) - \mathrm{Tr}(O^{\otimes 2}\,\mathbb{E}[\hat{M}_1]^{\otimes 2}) \\ &= \mathrm{Tr}\Big(O^{\otimes 2}\,\mathrm{Var}[\hat{M}_1]\Big).\end{aligned}$$

7. Often this estimator would be called $\hat{\rho}$, and later we will rename it $\hat{\phi}$, but within this section we will use $\hat{M}_1$ since it is an unbiased estimator for M_1.
8. It will be important later how *often* failure occurs, and this is explored in Theorem 33.

We therefore take the trace with $O^{\otimes 2}$ and make a few simplifications – dropping negative terms, rounding $\frac{d+n}{d+n+1}$ up to 1, and so on.

$$\begin{aligned}
&\operatorname{Var}[\operatorname{Tr}(O\hat{M}_1)] \\
&= \operatorname{Tr}\left(O^{\otimes 2}\tfrac{d+n}{d+n+1}\left(\frac{W_{(1\,2)}(\mathbb{I}+nM_1\otimes\mathbb{I}+n\mathbb{I}\otimes M_1)}{n^2}+\frac{n-1}{n}M_2-M_1^{\otimes 2}-\frac{(\mathbb{I}+nM_1)^{\otimes 2}}{n^2(d+n)^2}\right)\right) \\
&= \tfrac{d+n}{d+n+1}\left(\frac{\operatorname{Tr}(O^2)}{n^2}+\frac{2\operatorname{Tr}(O^2M_1)}{n}+\frac{n-1}{n}\operatorname{Tr}(O^{\otimes 2}M_2)-\operatorname{Tr}(OM_1)^2-\left(\frac{\operatorname{Tr}(O)+n\operatorname{Tr}(OM_1)}{n(d+n)}\right)^2\right) \\
&\leq \frac{\operatorname{Tr}(O^2)}{n^2}+\frac{2\operatorname{Tr}(O^2M_1)}{n}+\frac{n-1}{n}\operatorname{Tr}(O^{\otimes 2}M_2)-\operatorname{Tr}(OM_1)^2.
\end{aligned}$$

Last, we apply Hölder's inequality for Schatten norms to bound $\operatorname{Tr}(O^2M_1) \leq \|O\|_\infty^2\|M_1\|_1 = \|O\|_\infty^2$, using the fact that $\|M_1\|_1 = 1$ because M_1 is a state (Lemma 23). ■

B.2.1. Pure State Classical Shadows

As a quick exercise, we can re-derive the mean and variance of the pure state classical shadows estimator from Grier et al. (2022). When ρ is pure, it is not hard to see that $\operatorname{Tr}_{n\to k}(\rho^{\otimes n}) = \rho^{\otimes k}$ and thus

$$M_1 = \frac{\operatorname{Tr}_{n\to 1}(\rho^{\otimes n})}{\operatorname{Tr}_{n\to 0}(\rho^{\otimes n})} = \frac{\rho}{1} = \rho, \qquad M_2 = \frac{\operatorname{Tr}_{n\to 2}(\rho^{\otimes n})}{\operatorname{Tr}_{n\to 0}(\rho^{\otimes n})} = \frac{\rho^{\otimes 2}}{1} = \rho^{\otimes 2}.$$

Thus $\hat{M}_1$ is an unbiased estimator for $M_1 = \rho$ (which is generally *not* the case when ρ is mixed).

Lemma 25 *Let O be a Hermitian observable. When ρ is pure, $\mathbb{E}[\operatorname{Tr}(O\hat{M}_1)] = \operatorname{Tr}(O\rho)$ and*

$$\operatorname*{Var}_{\text{meas}}[\operatorname{Tr}(O\hat{M}_1)] \leq \frac{\operatorname{Tr}(O^2)}{n^2}+\frac{2\|O\|_\infty^2}{n}.$$

Proof Use Corollary 24. The expectation of $\operatorname{Tr}(O\hat{M}_1)$ follows immediately. For the variance, we observe that $\operatorname{Tr}(O^{\otimes 2}M_2) = \operatorname{Tr}(O^{\otimes 2}\rho^{\otimes 2}) = \operatorname{Tr}(O\rho)^2$, which is then more than cancelled out by $\operatorname{Tr}(OM_1)^2 = \operatorname{Tr}(O\rho)^2$. ■

B.3. Classical mixture of orthogonal tensor products

Corollary 24 reduces the mean and variance calculation to computing M_1 and M_2. However, even for $n = 4$, $\operatorname{Tr}_{4\to 1}(\rho^{\otimes 4})$ is the unwieldy polynomial

$$\frac{1}{24}\big(\rho(1+3\operatorname{Tr}(\rho^2)+2\operatorname{Tr}(\rho^3))+\rho^2(3+\operatorname{Tr}(\rho^2))+6\rho^3+6\rho^4\big),$$

and $\operatorname{Tr}_{4\to 2}(\rho^{\otimes 4})$ is even worse. It is hard to say much about, e.g., the overlap of M_1 with the leading eigenvector (Φ_1), beyond the fact that it is some symmetric polynomial in the

eigenvalues of ρ. It is unclear how to bound it in terms of the deviation, $\eta = 1 - \lambda_1$, especially for arbitrary n. We need a different approach to compute M_1 and M_2 for $\rho^{\otimes n}$.

Recall that $\rho = \sum_{i=1}^{d} \lambda_i \Phi_i$ where $\Phi_i := |\phi_i\rangle\langle\phi_i|$ are projectors onto the eigenvectors, $|\phi_i\rangle$, which form an orthonormal basis. It follows that we can expand $\rho^{\otimes n}$ in a basis of n-fold tensor products, $\Phi_{\mathbf{a}} := \otimes_{i=1}^{n} \Phi_{a_i}$, where $\mathbf{a} = (a_1, \ldots, a_n) \in [d]^n$. That is,

$$\rho^{\otimes n} = \sum_{\mathbf{a} \in [d]} \left(\prod_{i=1}^{n} \lambda_{a_i} \right) \Phi_{\mathbf{a}}$$

Since $\rho^{\otimes n}$ is exchangeable, we can express the right hand side as a sum of exchangeable operators. Specifically, we group the terms by type, where the *type* of $\Phi_{\mathbf{a}}$ is a vector $\mathbf{e} = (e_1, \ldots, e_d)$ where $e_i := \#\{j : a_j = i\}$ is the number of occurrences of Φ_i. The grouped states we call $\sigma(\mathbf{e})$.

Definition 26 *Fix a basis $|\phi_1\rangle, \ldots, |\phi_d\rangle$, and let $\Phi_i = |\phi_i\rangle\langle\phi_i|$ for all $i = 1, \ldots, d$. Given $\mathbf{e} = (e_1, \ldots, e_d) \in \mathbb{N}^d$ such that $e_1 + \cdots + e_d = n$, define a mixed state*

$$\sigma(e_1, \ldots, e_d) := \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} W_\pi \left(\bigotimes_{j=1}^{d} \Phi_j^{\otimes e_j} \right) W_\pi^\dagger \in \mathrm{Dens}\Big((\mathbb{C}^d)^{\otimes n} \Big).$$

Let us define some shorthand notation with $\mathbf{e}$ for later use.

$$\mathbf{e}! := e_1! \cdots e_d!, \qquad \binom{n}{\mathbf{e}} := \frac{n!}{\mathbf{e}!}, \qquad \underline{\lambda}^{\mathbf{e}} := \lambda_1^{e_1} \cdots \lambda_d^{e_d},$$

We can now succinctly write $\rho^{\otimes n}$ as a convex combination of $\sigma(\mathbf{e})$.

Proposition 27

$$\rho^{\otimes n} = \sum_{\mathbf{e}} \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \sigma(\mathbf{e})$$

A mixture of this form is indistinguishable from a $\sigma(\mathbf{e})$ selected at random from the distribution with mass $\Pr[\mathbf{e}] = \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}}$, which we recognize as a multinomial distribution with n trials for d events with probabilities $\lambda_1, \ldots, \lambda_d$. However, the pertinent distribution for our calculations, which we name $\mathscr{D}$, is this multinomial conditioned on successful measurement, since $\hat{M}_1$ is conditional on a successful measurement outcome. That is, the probability mass function for for $\mathscr{D}$ is

$$\Pr[\mathbf{e} \mid \text{success}] = \frac{\Pr[\text{success} \mid \mathbf{e}] \Pr[\mathbf{e}]}{\Pr[\text{success}]} \propto \Pr[\text{success} \mid \mathbf{e}] \cdot \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}}.$$

We will calculate shortly (Theorem 29), the probability of success for a given $\mathbf{e}$, to make this distribution more explicit.

Recall that the purpose of decomposing $\rho^{\otimes n}$ as a mixture (Proposition 27) was to give another path to compute M_1 and M_2. To this end, we define

$$M_k^{\mathbf{e}} := \frac{\mathrm{Tr}_{n \to k}(\sigma(\mathbf{e}))}{\mathrm{Tr}_{n \to 0}(\sigma(\mathbf{e}))},$$

for all $k \geq 1$, and prove the following connection to the original M_ks.

Theorem 28 *Fix $0 \le k \le n$, and then $M_k(\rho^{\otimes n}) = \mathbb{E}_{\mathbf{e}\sim\mathscr{D}}[M_k(\sigma(\mathbf{e}))]$, i.e., $M_k(\rho^{\otimes n})$ is the expectation of $M_k(\sigma(\mathbf{e}))$ over $\mathbf{e}$ sampled from $\mathscr{D}$.*

Proof

$$\begin{aligned}
M_k(\rho^{\otimes n}) &= \frac{\mathrm{Tr}_{n\to k}(\rho^{\otimes n})}{\mathrm{Tr}_{n\to 0}(\rho^{\otimes n})} && \text{definition}\\
&= \frac{\mathrm{Tr}_{n\to k}(\rho^{\otimes n})}{\Pr[\text{success}]} && \text{because } \Pr[\text{success}] = \operatorname*{Tr}_{n\to 0}(\rho^{\otimes n})\\
&= \frac{1}{\Pr[\text{success}]}\sum_{\mathbf{e}} \Pr[\mathbf{e}]\cdot \operatorname*{Tr}_{n\to k}(\sigma(\mathbf{e})) && \text{linearity of trace}\\
&= \sum_{\mathbf{e}} \frac{\Pr[\mathbf{e}]}{\Pr[\text{success}]}\cdot \operatorname*{Tr}_{n\to k}(\sigma(\mathbf{e}))\cdot \frac{\Pr[\text{success}\mid \mathbf{e}]}{\mathrm{Tr}_{n\to 0}(\sigma(\mathbf{e}))} && \text{since } \Pr[\text{success}\mid \mathbf{e}] = \operatorname*{Tr}_{n\to 0}(\sigma(\mathbf{e}))\\
&= \sum_{\mathbf{e}} \frac{\Pr[\text{success}\mid \mathbf{e}]\Pr[\mathbf{e}]}{\Pr[\text{success}]}\cdot \frac{\mathrm{Tr}_{n\to k}(\sigma(\mathbf{e}))}{\mathrm{Tr}_{n\to 0}(\sigma(\mathbf{e}))} && \text{rearrange}\\
&= \sum_{\mathbf{e}} \Pr[\mathbf{e}\mid \text{success}]\cdot M_k(\sigma(\mathbf{e})) && \text{Bayes' rule}\\
&= \mathbb{E}[M_k(\sigma(\mathbf{e}))\mid \text{success}].
\end{aligned}$$

■

That is, $M_k = \mathbb{E}[M_k^{\mathbf{e}}]$. We stress that there are now two sources of randomness affecting our estimator: *mixture randomness* arising from a random choice of initial state $\sigma(\mathbf{e})$, and *measurement randomness* caused by the inherent randomness of measuring a quantum state.

B.4. M_1 and M_2 for $\sigma(\mathbf{e})$

Section B.1 and Section B.2 showed that moments of Ψ and $\hat{M}_1$ can be calculated from M_1 and M_2, then Section B.3 expressed M_1 and M_2 as expectations of $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$. In this section, we calculate

$$Z^{\mathbf{e}} = \Pr[\text{success}\mid \mathbf{e}] = \operatorname*{Tr}_{n\to 0}(\sigma(\mathbf{e})),$$

as well as $M_1^{\mathbf{e}}$, and $M_2^{\mathbf{e}}$.

As a starting point for all three calculations, we have that

$$\operatorname*{Tr}_{n\to k}(\sigma(\mathbf{e})) = \operatorname*{Tr}_{[n-k]}(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e})) = \operatorname*{Tr}_{[n-k]}(\Pi_{\mathrm{sym}}^{(n)}\Phi^{\mathbf{e}}) = \frac{1}{n!}\sum_{\pi\in \mathrm{S}_n}\operatorname*{Tr}_{[n-k]}(W_\pi\Phi^{\mathbf{e}})$$

where we define $\Phi^{\otimes\mathbf{e}} = \bigotimes_{i=1}^{d}\Phi_j^{\otimes e_j}$.

We claim that expressions like $\mathrm{Tr}_{[n-k]}(W_\pi\Phi^{\otimes\mathbf{e}})$ can be evaluated combinatorially — the permutation will loop through various Φ_i tensor factors of $\Phi^{\otimes\mathbf{e}}$, and if any two adjacent Φ_i and Φ_j are orthogonal (we call this *mixing eigenstates* in the theorems that follow), then the entire term vanishes. If not, then many of the Φ_i are traced out, and those that remain appear in the result.

The following proofs may be a bit opaque if the reader cannot visualize $\mathrm{Tr}_{[n-k]}(W_\pi\Phi^{\otimes\mathbf{e}})$ in the language of tensor networks. Tensor networks are a graphical model of linear operators

where, e.g., the permutation operator W_π is drawn as a literal permutation of legs, and partial trace is achieved by looping the output legs for the traced-out qudits back to the input legs. We cannot include a full introduction to tensor networks here, but refer the reader to either Roberts and Yoshida (2017) or Grier et al. (2022) to see examples of these ideas in action on very similar problems.

With that, we begin by calculating the full trace. Since $\Pi_{\mathrm{sym}}^{(n)}$ is the sum of the successful measurement outcomes, it also represents the probability of success.

Theorem 29 *For all $n \geq 1$ and $\mathbf{e}$ such that $e_1 + \cdots + e_d = n$,*

$$Z^{\mathbf{e}} := \operatorname*{Tr}_{n\to 0}(\sigma(\mathbf{e})) = \operatorname{Tr}(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e})) = \binom{n}{\mathbf{e}}^{-1}.$$

Proof Expand σ with the definition and $\Pi_{\mathrm{sym}}^{(n)}$ as an average over permutations:

$$\operatorname{Tr}\Big(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e})\Big) = \frac{1}{(n!)^2}\sum_{\pi,\sigma\in \mathrm{S}_n} \operatorname{Tr}\left(W_\pi\left(\bigotimes_{j=1}^{d}\Phi_j^{\otimes e_j}\right)W_\pi^\dagger W_\sigma\right) = \frac{1}{n!}\sum_{\pi\in\mathrm{S}_n}\operatorname{Tr}\big(\Phi_1^{\otimes e_1}\cdots\Phi_d^{\otimes e_d}W_\pi\big).$$

Any permutation π which *mixes eigenstates* by having a cycle which involves Φ_i and Φ_j for $i \neq j$ may be ignored, since the trace factor for that cycle will be zero. If a permutation does not mix eigenstates, then we have $\operatorname{Tr}\big(\Phi_1^{\otimes e_1}\cdots\Phi_d^{\otimes e_d}W_\pi\big) = 1$, since it is a product of traces of the form $\operatorname{Tr}(\Phi_i^k)$, each of which is 1 because Φ_i is pure.

In other words, it suffices to count permutations which do not mix eigenstates. Clearly there are $e_i!$ ways to permute the $\Phi_i^{\otimes e_i}$ factors among themselves, and we make this choice independently for each i for a total of $\mathbf{e}!$ permutations which give unit trace. Therefore,

$$\operatorname{Tr}\Big(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e})\Big) = \frac{\mathbf{e}!}{n!} = \binom{n}{\mathbf{e}}^{-1},$$

completing the proof. ∎

Next, we find that $M_1^{\mathbf{e}}$ has a surprisingly clean form.

Theorem 30 *For all $n \geq 1$ and $\mathbf{e}$ such that $e_1 + \cdots + e_d = n$,*

$$M_1^{\mathbf{e}} := \frac{\operatorname{Tr}_{n\to 1}(\sigma(\mathbf{e}))}{\operatorname{Tr}_{n\to 0}(\sigma(\mathbf{e}))} = \frac{\operatorname{Tr}_{1,\ldots,n-1}(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e}))}{Z^{\mathbf{e}}} = \frac{1}{n}\sum_{i=1}^{d} e_i\Phi_i.$$

Proof For any exchangeable $A \in \operatorname{Dens}\big((\mathbb{C}^d)^{\otimes n}\big)$, we have

$$\operatorname*{Tr}_{[n-1]}(A) = \operatorname*{Tr}_{-1}(A) = \operatorname*{Tr}_{-i}(W_\pi A W_{\pi^{-1}}) = \operatorname*{Tr}_{-i}(A)$$

for any $\pi \in \mathrm{S}_n$ such that $\pi(i) = n$. Therefore, we can expand $\Pi_{\mathrm{sym}}^{(n)}$ as an average over permutations, and also average over the indices which are traced out to obtain

$$\operatorname*{Tr}_{n\to 1}(\sigma(\mathbf{e})) = \operatorname*{Tr}_{[n-1]}(\Pi_{\mathrm{sym}}^{(n)}\sigma(\mathbf{e})) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n!}\sum_{\pi\in\mathrm{S}_n}\operatorname*{Tr}_{-i}\left(W_\pi\bigotimes_{j=1}^{d}\Phi_j^{\otimes e_j}\right)$$

As before, if π mixes eigenstates Φ_i and Φ_j in the same cycle (for $i \neq j$), then that term contributes 0. There are $e_1! \cdots e_d!$ permutations which do not mix eigenstates. Suppose index i corresponds to Φ_j. In each of the non-mixing permutations, the partial trace is Φ_j. Since there are e_j indices where there is a Φ_j, it follows that

$$\operatorname*{Tr}_{n\to 1}(\sigma(\mathbf{e})) = \binom{n}{\mathbf{e}}^{-1} \frac{1}{n} \sum_{i=1}^{d} e_i \Phi_i.$$

Dividing through by $Z^{\mathbf{e}} = \operatorname{Tr}_{n\to 0}(\sigma(\mathbf{e})) = \binom{n}{\mathbf{e}}^{-1}$ finishes the proof. ■

Theorem 31 *For all $n \geq 1$ and $\mathbf{e}$ such that $e_1 + \cdots + e_d = n$,*

$$M_2^{\mathbf{e}} = \frac{\operatorname{Tr}_{n\to 1}(\sigma(\mathbf{e}))}{\operatorname{Tr}_{n\to 0}(\sigma(\mathbf{e}))} = \frac{\operatorname{Tr}_{1,\ldots,n-1}(\Pi_{\mathrm{sym}}^{(n)} \sigma(\mathbf{e}))}{Z^{\mathbf{e}}} = \frac{2\Pi_{\mathrm{sym}}^{(2)}}{n(n-1)} \left(\sum_{i \neq j} e_i e_j \Phi_i \otimes \Phi_j + \sum_k \binom{e_k}{2} \Phi_k^{\otimes 2} \right).$$

Proof Expand $\Pi_{\mathrm{sym}}^{(n)}$ as an average over permutations π, and average over an ordered pair of distinct indices, a and b.

$$\operatorname*{Tr}_{[n-2]}(\Pi_{\mathrm{sym}}^{(n)} \sigma(\mathbf{e})) = \frac{1}{n(n-1)} \sum_{a \neq b} \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} \operatorname*{Tr}_{-a,-b} \left(W_\pi \bigotimes_{k=1}^{d} \Phi_k^{\otimes e_k} \right)$$

For each a, b, divide the permutations into those where a, b are in separate cycles (type I), and those where a, b are in the same cycle (type II). We note that composing a permutation with the transposition $(a\, b)$ also changes the type, and since this operation is clearly invertible, it is a bijection between the two types of permutations. That is, there are equally many type I and type II permutations. In fact, we can pair up the permutations π, π' (one of each type) matched by the bijection, factor $(W_{(a)(b)} + W_{(a\,b)})$ out of $W_\pi + W_{\pi'}$ and out of the partial trace where it becomes $2\Pi_{\mathrm{sym}}^{(2)}$. In other words, it suffices to analyze type I permutations and then multiply by $2\Pi_{\mathrm{sym}}^{(2)}$.

Let us say a position a is *incident* to state Φ_i if Φ_i is the ath term of the tensor product. The first case we are interested in is when a and b are incident to the same state, Φ_i. We observe that there are $e_1! \cdots e_d!$ non-mixing permutations total, but this includes both type I and type II permutations. There are $e_i(e_i - 1)$ ways to pick a and b from $\Phi_i^{\otimes e_i}$, but we divide this in half to get those of type I. Hence, when a and b are incident to the same Φ_i, we have

$$\sum_{\pi \in \text{Type I}} \operatorname*{Tr}_{-a,-b} \left(W_\pi \bigotimes_{k=1}^{d} \Phi_k^{\otimes e_k} \right) = \mathbf{e}! \binom{e_i}{2} \Phi_i \otimes \Phi_i.$$

The second case is a incident to Φ_i and b incident to Φ_j for $i \neq j$. There are $e_1! \cdots e_d!$ type I permutations which do not mix eigenstates, and thus have a nonzero partial trace to contribute to the sum. Note that $\mathbf{e}!$ over $n!$ from the average over permutations gives the

$\binom{n}{\mathbf{e}}^{-1}$ we've come to expect in these calculations. There are e_i indices a within $\Phi_i^{\otimes e_i}$, and e_j indices b within $\Phi_j^{\otimes e_j}$. To summarize, we get

$$\sum_{\pi \in \text{Type I}} \operatorname*{Tr}_{-a,-b}\left(W_\pi \bigotimes_{k=1}^{d} \Phi_k^{\otimes e_k}\right) = \mathbf{e}! \cdot e_i e_j \Phi_i \otimes \Phi_j.$$

Altogether, this leads to

$$\begin{aligned}
\operatorname*{Tr}_{[n-2]}(\Pi_{\text{sym}}^{(n)} \sigma(\mathbf{e})) &= \frac{1}{n(n-1)} \sum_{a \neq b} \frac{1}{n!} \sum_{\pi \in \mathrm{S}_n} \operatorname*{Tr}_{-a,-b}\left(W_\pi \bigotimes_{k=1}^{d} \Phi_k^{\otimes e_k}\right) \\
&= \frac{2\Pi_{\text{sym}}^{(2)}}{n(n-1)} \sum_{a \neq b} \frac{1}{n!} \sum_{\pi \in \text{Type I}} \operatorname*{Tr}_{-a,-b}\left(W_\pi \bigotimes_{k=1}^{d} \Phi_k^{\otimes e_k}\right) \\
&= \frac{2\Pi_{\text{sym}}^{(2)}}{n(n-1)} \cdot \frac{\mathbf{e}!}{n!}\left(\sum_i \binom{e_i}{2} \Phi_i \otimes \Phi_i + \sum_{i \neq j} e_i e_j \Phi_i \otimes \Phi_j\right),
\end{aligned}$$

and dividing by $Z^{\mathbf{e}} = \operatorname{Tr}_{n \to 0}(\sigma(\mathbf{e})) = \binom{n}{\mathbf{e}}^{-1}$ yields the stated result. ■

Let us combine and summarize Theorems 29, 30, and 31 in one result.

Corollary 32 *For all $n \geq 1$,*

$$Z := \operatorname{Tr}\left(\Pi_{\text{sym}}^{(n)} \rho^{\otimes n}\right) = \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}},$$

$$M_1 = \frac{1}{Z} \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \frac{1}{n}\left(\sum_{i=1}^{d} e_i \Phi_i\right),$$

$$M_2 = \frac{1}{Z} \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \frac{2\Pi_{\text{sym}}^{(2)}}{n(n-1)}\left(\sum_{i \neq j} e_i e_j \Phi_i \otimes \Phi_j + \frac{1}{2} \sum_k e_k(e_k - 1) \Phi_k \otimes \Phi_k\right),$$

where all three outer sums are over $\mathbf{e} \in \mathbb{N}^d$ such that $e_1 + \cdots + e_d = n$.

Proof For all k we have,

$$\operatorname*{Tr}_{n \to k}(\rho^{\otimes n}) = \sum_{\mathbf{e}} \binom{n}{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \operatorname*{Tr}_{n \to k}(\sigma(\mathbf{e})) = \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} \frac{\operatorname{Tr}_{n \to k}(\sigma(\mathbf{e}))}{\operatorname{Tr}_{n \to 0}(\sigma(\mathbf{e}))} = \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}} M_k^{\mathbf{e}},$$

and then dividing through by Z gives the results for M_1 and M_2. ■

Since $Z^{\mathbf{e}} = \Pr[\text{success} \mid \mathbf{e}] = \binom{n}{\mathbf{e}}^{-1}$, and $Z = \Pr[\text{success}]$ we can now see that the probability mass function for $\mathscr{D}$ is

$$\Pr[\mathbf{e} \mid \text{success}] = \frac{\Pr[\text{success} \mid \mathbf{e}] \Pr[\mathbf{e}]}{\Pr[\text{success}]} = \frac{\underline{\lambda}^{\mathbf{e}}}{Z}.$$

In light of this, the expressions in the corollary appear to be expectations over $\mathbf{e} \sim \mathscr{D}$, which is precisely what Theorem 28 proves, so everything squares up nicely. It is a good time to also bound the probability the measurement is successful in terms of λ_1.

Theorem 33 *The probability of a successful measurement is Z and*

$$\lambda_1^{n-1} \leq Z \leq \frac{\lambda_1^{n+1}}{2\lambda_1 - 1} = \lambda_1^{n-1}(1 + \mathcal{O}(\eta^2)) \tag{14}$$

Proof Recall that Z is the probability of success and $Z = \sum_{\mathbf{e}} \underline{\lambda}^{\mathbf{e}}$ over $\mathbf{e}$ totaling n. On the one hand, it is lower bounded by the terms where $e_1 = n$ (i.e., λ_1^n) and $e_1 = n-1$ (i.e., $\lambda_1^{n-1}(\lambda_2 + \cdots + \lambda_d)$). It follows that

$$Z \geq \lambda_1^n + \lambda_1^{n-1}(\lambda_2 + \cdots + \lambda_d) = \lambda_1^{n-1}(\lambda_1 + \cdots + \lambda_d) = \lambda^{n-1}.$$

On the other hand, since λ_1 is the dominant eigenvalue, it also makes sense to expand around it.

$$Z = \sum_{e_1 + \cdots + e_d = n} \underline{\lambda}^{\mathbf{e}} = \sum_{e_1=0}^{n} \sum_{e_2 + \cdots + e_d = n - e_1} \underline{\lambda}^{\mathbf{e}}.$$

We can then insert multinomial coefficients to simplify the $\lambda_2, \ldots, \lambda_d$ part.

$$Z \leq \sum_{e_1=0}^{n} \sum_{e_2 + \cdots + e_d = n - e_1} \binom{n - e_1}{e_2, \ldots, e_d} \underline{\lambda}^{\mathbf{e}} = \sum_{e_1=0}^{n} \lambda_1^{e_1}(\lambda_2 + \cdots + \lambda_d)^{n - e_1}$$

Re-indexing and letting the sum extend to infinity, we have

$$Z \leq \sum_{j=0}^{n} \lambda_1^{n-j}(1 - \lambda_1)^j = \lambda_1^n \sum_{j=0}^{\infty} \left(\frac{1 - \lambda_1}{\lambda_1}\right)^j = \frac{\lambda_1^{n+1}}{2\lambda_1 - 1}.$$

Finally, $\eta = 1 - \lambda_1$ and we note that

$$\frac{\lambda_1^2}{2\lambda_1 - 1} = \frac{(1 - \eta)^2}{1 - 2\eta} = 1 + \eta^2 + \mathcal{O}(\eta^3),$$

so the (multiplicative) gap between the two bounds is only $1 + \eta^2 + \mathcal{O}(\eta^3)$. ■

B.5. Geometric approximation

We now have expressions for the mean and variance of $\hat{M}_1$ in terms of M_1 and M_2 (Corollary 24), expressions for M_1 and M_2 as expectations over $M_1^{\mathbf{e}}$, $M_2^{\mathbf{e}}$ (Theorem 28), expressions for $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ in terms of $\mathbf{e}$ (Theorem 30, Theorem 31), and the distribution $\mathscr{D}$ for the expectation. There is one last obstacle to overcome: we would like to compute $\mathbb{E}_{\mathbf{e} \in \mathscr{D}}[e_i]$ and $\mathbb{E}_{\mathbf{e} \in \mathscr{D}}[e_i e_j]$, since those appear in M_1 and M_2. Exact expressions for these expectations have eluded us,[9] so we define an approximation, $\mathscr{D}'$, of the true distribution such that $\mathbb{E}_{\mathbf{e} \in \mathscr{D}'}[e_i]$ is straightforward.

Suppose the first eigenvalue is much larger than the rest, i.e., $\lambda_1 \gg \lambda_2 \geq \cdots \geq \lambda_d$. Hence, the $\mathbf{e}$ vectors with highest probability in $\mathscr{D}$ have e_1 close to n, as large as possible.

9. Also Z, which we could only upper and lower bound in Theorem 33.

Let us rewrite the probability mass using the fact that $e_1 = n - e_2 - \cdots - e_d$. As long as $e_1, \ldots, e_d \in \mathbb{N}$, we have

$$f(\mathbf{e}) = \frac{1}{Z}\lambda_1^{n-e_2-\ldots-e_d}\prod_{i=2}^{d}\lambda_i^{e_i} = \frac{\lambda_1^n}{Z}\prod_{i=2}^{d}\left(\frac{\lambda_i}{\lambda_1}\right)^{e_i}.$$

It *appears* that $\mathscr{D}$ factors as a product distribution on $e_2, \ldots, e_d$, i.e., it is proportional to $f_2(e_2)\cdots f_d(e_d)$ where $f_i(e_i) = (\frac{\lambda_i}{\lambda_1})^{e_i}(1 - \frac{\lambda_i}{\lambda_1})$ is the p.d.f. of a geometric random variable with mean $\frac{\lambda_i}{\lambda_1 - \lambda_i}$. We know $e_2, \ldots, e_d$ are not independent in $\mathscr{D}$, so there is a catch: in the *very* unlikely event that $e_2 + \ldots + e_d$ exceeds n, the condition $e_1 = n - (e_2 + \cdots + e_d)$ requires us to set $e_1 < 0$. In fact, this is the only difference between the distributions.

Lemma 34 *The distribution $\mathscr{D}$ is exactly $\mathscr{D}'$ conditioned on $e_1 \geq 0$.*

Proof For $\mathbf{e}$ with $e_1 \geq 0$ (the full support of $\mathscr{D}$) we have already seen that the p.d.f. f factors as a product of f_i (times a constant).

$$f(\mathbf{e}) = \frac{1}{Z}\lambda_1^{n-e_2-\ldots-e_d}\prod_{i=2}^{d}\lambda_i^{e_i} = \frac{\lambda_1^n}{Z}\prod_{i=2}^{d}\left(\frac{\lambda_i}{\lambda_1}\right)^{e_i} = \frac{\lambda_1^n}{Z}\prod_{i=2}^{d}\frac{f_i(e_i)}{1-\frac{\lambda_i}{\lambda_1}}$$

That is, whenever $e_1, \ldots, e_d \geq 0$, the two distributions are proportional. The only other $\mathbf{e}$ with any support in $\mathscr{D}'$ are those with $e_1 < 0$, therefore if we condition on $e_1 \geq 0$ then $\mathscr{D}'$ becomes $\mathscr{D}$. ■

The two distributions are *very* close to each other. We have consigned the proofs to Appendix D, but we quote the highlights below. First, the probability that $e_1 < 0$ is indeed very small, which in turn bounds the total variation distance, $\|\mathscr{D} - \mathscr{D}'\|_{TV}$.

Theorem 35

$$\Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 < 0] \leq \Delta := \left(\frac{1-\lambda_1}{\lambda_1}\right)^{n+1}\frac{\lambda_1}{2\lambda_1 - 1}.$$

It follows that $\|\mathscr{D} - \mathscr{D}'\|_{TV} = \Delta$.

When we rewrite Δ in terms of $\eta \leq \frac{1}{3}$,

$$\Delta = \left(\frac{\eta}{1-\eta}\right)^{n+1}\frac{1-\eta}{1-2\eta} \leq 2\cdot\left(\tfrac{3}{2}\eta\right)^{n+1},$$

we see that $\Delta = \mathcal{O}(\eta^2)$, even if $n = 1$. More realistically, we will have $n \approx \frac{1}{\eta}$, and then Δ vanishes even more quickly, as shown in the plot Figure 2.

We separately bound the change in $\mathbb{E}[\mathbf{e}]$ for $\mathscr{D}$ versus $\mathscr{D}'$.

Theorem 36 *The difference between $\mathscr{D}$ and $\mathscr{D}'$ for first-order expectations is at most*

$$\left\|\mathop{\mathbb{E}}_{\mathscr{D}}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}]\right\|_1 = \sum_i\left|\mathop{\mathbb{E}}_{\mathscr{D}}[e_i] - \mathop{\mathbb{E}}_{\mathscr{D}'}[e_i]\right| \leq \frac{2\Delta}{1-\Delta}\left(n + \frac{1}{2\lambda_1 - 1}\right).$$

Last, the variance difference is quantified with covariance matrices.

Lemma 37 *The covariance matrices of $\mathscr{D}$ and $\mathscr{D}'$ are related as follows:*

$$\Sigma_{\mathscr{D}'} \succeq \Sigma_{\mathscr{D}}(1-\Delta)^2.$$

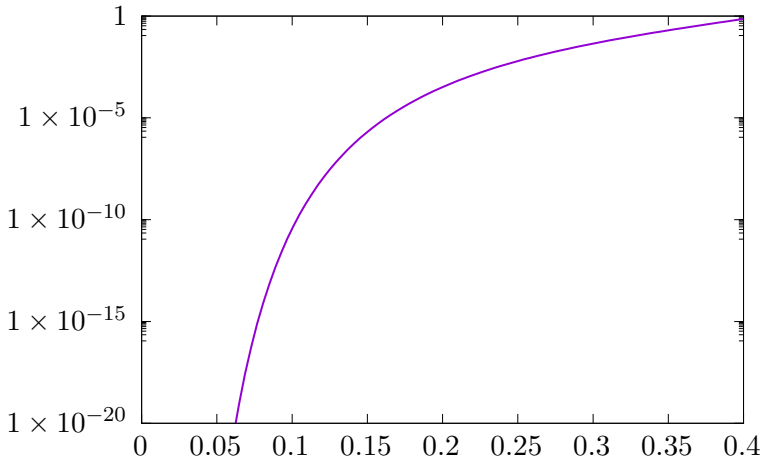


Figure 2: An upper bound on the mass of $\mathscr{D}'$ outside the support of $\mathscr{D}$ (as in Theorem 35) as a function of $\eta := 1 - \lambda_1$, assuming $n = 1/\eta$, semilog scale.

B.6. Mean of the estimator

In this section, we find the mean of the estimator $\hat{M}_1$ conditioned on the success of the measurement (on state ρ). Recall that $\mathbb{E}[\hat{M}_1] = M_1$ by Corollary 24, and Theorem 28 expands this into

$$M_1 := M_1(\rho^{\otimes n}) = \mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[M_1(\sigma(\mathbf{e}))] = \mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[M_1^{\mathbf{e}}].$$

Below we approximate this expectation, except with the geometric random variable distribution $\mathscr{D}'$ in place of $\mathscr{D}$

Theorem 38 (Mean with $\mathscr{D}'$)

$$\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}[M_1^{\mathbf{e}}] = \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}\left[\frac{1}{n}\left(\sum_{i=1}^{d} e_i\Phi_i\right)\right] = \Phi_1 + \frac{1}{n}\sum_{j=2}^{d}\tfrac{\lambda_j}{\lambda_1-\lambda_j}(\Phi_j - \Phi_1)$$

Proof We rewrite with $e_1 = n - e_2 - \cdots - e_d$ and use that $\mathbb{E}_{\mathbf{e}\in\mathscr{D}'}[e_i] = \frac{\lambda_i}{\lambda_1-\lambda_i}$ for all $2 \le i \le d$ (by Proposition 50).

$$\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}\left[\frac{1}{n}\left(\sum_{i=1}^{d} e_i\Phi_i\right)\right] = \frac{1}{n}\left(n\Phi_1 + \sum_{i=2}^{d}\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}[e_i](\Phi_i - \Phi_1)\right) = \Phi_1 + \frac{1}{n}\sum_{i=2}^{d}\frac{\lambda_i}{\lambda_1-\lambda_i}(\Phi_i - \Phi_1).$$

■

Corollary 39

$$\left\| M_1 - \Phi_1 - \frac{1}{n}\sum_{j=2}^{d}\tfrac{\lambda_j}{\lambda_1-\lambda_j}(\Phi_j - \Phi_1)\right\|_{\mathrm{tr}} = \mathcal{O}(\eta^2)$$

Proof This result is about comparing $M_1 = \mathbb{E}_{\mathbf{e}\in\mathscr{D}}[M_1^{\mathbf{e}}]$ with

$$\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}[M_1^{\mathbf{e}}] = \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}\left[\frac{1}{n}\left(\sum_{i=1}^{d} e_i\Phi_i\right)\right] = \Phi_1 + \frac{1}{n}\sum_{j=2}^{d}\tfrac{\lambda_j}{\lambda_1-\lambda_j}(\Phi_j - \Phi_1)$$

from the previous theorem. That is, we are bounding the difference in M_1 due to our approximation of $\mathscr{D}$ with $\mathscr{D}'$. Since the operators are diagonal in the Φ_i basis, the trace norm simplifies:

$$\frac{1}{n}\left(\left\|\sum_{i=1}^{d}(\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_i] - \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}[e_i])\Phi_i\right\|_{\mathrm{tr}}\right) = \frac{1}{n}\sum_{i=1}^{d}\left|\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_i] - \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}'}[e_i]\right|.$$

Theorem 36 upper bounds this by

$$\frac{1}{n}\frac{2\Delta}{1-\Delta}\left(n + \frac{1}{2\lambda_1 - 1}\right) = \mathcal{O}(\Delta) \subseteq \mathcal{O}(\eta^2).$$

■

B.7. Variance of the estimator

Recall that there are two sources of variance for $\mathrm{Tr}(O\hat{M}_1)$: mixture randomness (from $\rho^{\otimes n}$ being a mixture of pure states $\sigma(\mathbf{e})$), and the inherent randomness of quantum measurement. These two sources of randomness are responsible for variance of $\mathrm{Tr}(O\hat{M}_1)$, as formalized by the law of total variance.

Theorem 40 (Law of total variance)

$$\mathop{\mathrm{Var}}_{\mathbf{e}\sim\mathscr{D},\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1)) = \mathop{\mathrm{Var}}_{\mathbf{e}\sim\mathscr{D}}[\mathop{\mathbb{E}}_{\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1) \mid \mathbf{e})] + \mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[\mathop{\mathrm{Var}}_{\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1) \mid \mathbf{e})].$$

We go on to bound these terms individually. Since the expressions for variance get somewhat unwieldy, we introduce shorthand common terms involving O and Φ_i: let $O_i := \mathrm{Tr}(O\Phi_i)$ and $O_{ij} := \mathrm{Tr}(O\Phi_i O\Phi_j)$. In general $O_iO_j \neq O_{ij}$, but it will be important that

$$O_{kk} = \mathrm{Tr}(O\Phi_k O\Phi_k) = \mathrm{Tr}(O\Phi_k)^2 = O_k^2$$

for all $1 \leq k \leq d$. We also introduce $\hat{M}_1^{\mathbf{e}}$ to represent the estimator conditioned on $\sigma(\mathbf{e})$ being the input state.

B.7.1. Variance due to mixture randomness

Theorem 41 *The variance in* $\mathrm{Tr}(O\hat{M}_1)$ *due to* $\rho^{\otimes n}$ *being a mixture of* $\sigma(\mathbf{e})$ *is*

$$\mathop{\mathrm{Var}}_{\mathbf{e}\in\mathscr{D}}[\mathop{\mathbb{E}}_{\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1) \mid \mathbf{e})] = \mathop{\mathrm{Var}}_{\mathbf{e}\in\mathscr{D}}[\mathop{\mathbb{E}}_{\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1^{\mathbf{e}})] \leq \frac{4\|O\|_\infty^2}{n^2(1-\Delta)^2}\frac{\lambda_1(1-\lambda_1)}{(2\lambda_1-1)^2}.$$

Proof First, $\mathbb{E}_{\text{meas}}(\operatorname{Tr}(O\hat{M}_1^{\mathbf{e}}) \mid \mathbf{e}) = \operatorname{Tr}(O\,\mathbb{E}_{\text{meas}}(\hat{M}_1^{\mathbf{e}})) = \operatorname{Tr}(OM_1^{\mathbf{e}})$ by Corollary 24 — the corollary is for a general state A, so it applies with $\sigma(\mathbf{e})$.

Theorem 30 gives an expression for $M_1^{\mathbf{e}}$:

$$\operatorname{Tr}(OM_1^{\mathbf{e}}) = \frac{1}{n}\sum_{i=1}^{d} e_i \operatorname{Tr}(O\Phi_i) = \frac{1}{n}\sum_{i=1}^{d} e_i O_i = \frac{\mathbf{v}^\top \mathbf{e}}{n},$$

where $\mathbf{v} = (O_1, \ldots, O_d)$ is the vector of O_is. The variance is thus

$$\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}}[\operatorname{Tr}(OM_1^{\mathbf{e}})] = \frac{1}{n^2}\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}}[\mathbf{v}^\top \mathbf{e}] = \frac{\mathbf{v}^\top \Sigma_{\mathscr{D}} \mathbf{v}}{n^2},$$

where $\Sigma_{\mathscr{D}}$ is the covariance matrix for $\mathscr{D}$. By Lemma 37, $\Sigma_{\mathscr{D}} \preceq \frac{1}{(1-\Delta)^2}\Sigma_{\mathscr{D}'}$, and thus

$$\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}}[\operatorname{Tr}(OM_1^{\mathbf{e}})] = \frac{\mathbf{v}^\top \Sigma_{\mathscr{D}} \mathbf{v}}{n^2} \le \frac{\mathbf{v}^\top \Sigma_{\mathscr{D}'} \mathbf{v}}{n^2(1-\Delta)^2} = \frac{1}{(1-\Delta)^2}\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}'}\left[\frac{1}{n}\sum_{i=1}^{d} e_i O_i\right].$$

Since $e_2, \ldots, e_d$ are independent under $\mathscr{D}'$, we rewrite with $e_1 = n - e_2 - \cdots - e_d$ and simplify as much as possible.

$$\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}'}\left[\frac{1}{n}\sum_{i=1}^{d} e_i O_i\right] = \operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}'}\left[O_1 + \frac{1}{n}\sum_{i=2}^{d} e_i (O_i - O_1)\right] = \frac{1}{n^2}\sum_{i=2}^{d} \operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}'}[e_i](O_i - O_1)^2$$

Since $(O_i - O_1)^2 \le 4\|O\|_\infty^2$ and $\operatorname{Var}_{\mathbf{e}\sim\mathscr{D}'}[e_i] \le \frac{\lambda_1\lambda_i}{\lambda_1 - \lambda_i} \le \frac{\lambda_1\lambda_i}{2\lambda_1 - 1}$, the variance is bounded by

$$\operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}}[\operatorname{Tr}(OM_1^{\mathbf{e}})] \le \frac{1}{n^2(1-\Delta)^2}\sum_{i=2}^{d} \operatorname*{Var}_{\mathbf{e}\sim\mathscr{D}'}[e_i](O_i - O_1)^2 \le \frac{4\|O\|_\infty^2}{n^2(1-\Delta)^2}\frac{\lambda_1(1-\lambda_1)}{(2\lambda_1 - 1)^2}.$$

■

We note that the variance due to mixture randomness is small in all the ways we want: it is a function of $\|O\|^2$ rather than $\operatorname{Tr}(O^2)$, it is quadratic (rather than linear) in $\frac{1}{n}$, and it is multiplied by a factor of $\eta = 1 - \lambda_1$. We proceed with the analysis of the other (dominant) term in the variance.

B.7.2. Variance due to measurement randomness

Lemma 42

$$\operatorname*{Var}_{\text{meas}}[\operatorname{Tr}(O\hat{M}_1) \mid \mathbf{e}] \le \frac{\operatorname{Tr}(O^2)}{n^2} + \frac{2\|O\|_\infty^2}{n} + \frac{1}{n^2}\sum_{i\ne j} e_i e_j O_{ij} \tag{15}$$

Proof Recall that Corollary 24 already gives

$$\operatorname*{Var}_{\text{meas}}[\operatorname{Tr}(O\hat{M}_1^{\mathbf{e}})] \le \frac{\operatorname{Tr}(O^2)}{n^2} + \frac{2\|O\|_\infty^2}{n} + \frac{n-1}{n}\operatorname{Tr}(O^{\otimes 2} M_2^{\mathbf{e}}) - \operatorname{Tr}(OM_1^{\mathbf{e}})^2,$$

by applying it to $\sigma(\mathbf{e})$. The first two terms match the goal, so we focus on bounding the last two terms, using the expressions for $M_1^{\mathbf{e}}$ and $M_2^{\mathbf{e}}$ from Theorem 30 and Theorem 31.

$$\begin{aligned}
n(n-1)\operatorname{Tr}(O^{\otimes 2}M_2^{\mathbf{e}}) &= \operatorname{Tr}\left(O^{\otimes 2}2\Pi_{\mathrm{sym}}^{(2)}\left(\sum_{i\neq j} e_i e_j \Phi_i \otimes \Phi_j + \sum_k \binom{e_k}{2}\Phi_k^{\otimes 2}\right)\right) \\
&= \sum_{i\neq j} e_i e_j (O_i O_j + O_{ij}) + \tfrac{1}{2}\sum_k (e_k^2 - e_k)(O_k^2 + O_{kk}) \\
&= \sum_{i\neq j} e_i e_j O_i O_j + \sum_{i\neq j} e_i e_j O_{ij} + \sum_k e_k^2 O_k^2 - \sum_k e_k O_k^2 \\
&= \sum_{i,j} e_i e_j O_i O_j + \sum_{i\neq j} e_i e_j O_{ij} - \sum_k e_k O_k^2 \\
&= n^2 \operatorname{Tr}(OM_1^{\mathbf{e}})^2 + \sum_{i\neq j} e_i e_j O_{ij} - \sum_k e_k O_k^2
\end{aligned}$$

We can drop the negative term, and then it follows that

$$\frac{n-1}{n}\operatorname{Tr}(O^{\otimes 2}M_2^{\mathbf{e}}) - \operatorname{Tr}(OM_1^{\mathbf{e}})^2 \leq \sum_{i\neq j} e_i e_j O_{ij},$$

from which we get the result. ■

Now let us separately bound the last term of (15).

Lemma 43

$$\frac{1}{n^2}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}\left[\sum_{i\neq j} e_i e_j O_{ij}\right] \leq \frac{2}{n}\frac{1-\lambda_1}{2\lambda_1 - 1}\|O\|_\infty^2 + \mathcal{O}(\Delta)$$

Proof First, observe that

$$O_{ij} = \operatorname{Tr}(O\Phi_i O\Phi_j) = \langle\phi_i|O|\phi_j\rangle\langle\phi_j|O|\phi_i\rangle = |\langle\phi_i|O|\phi_j\rangle|^2 \leq \|O\|_\infty^2.$$

In other words, we can bound each O_{ij} by $\|O\|_\infty^2$, and then it clearly suffices to bound $\sum_{i\neq j} e_i e_j$. To start, we can look at the sum as being over all $e_i e_j$ – which totals n^2 on the basis that $e_1 + \cdots + e_d = n$ for all $\mathbf{e}$ – minus the "diagonal" terms, of which we claim only e_1^2 will be relevant.

$$\begin{aligned}
\sum_{i\neq j}\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_i e_j] &= \sum_{i,j}\mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_i e_j] - \sum_k \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_k^2] \\
&= \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}\left[\sum_{i,j} e_i e_j\right] - \sum_k \mathop{\mathbb{E}}_{\mathbf{e}\in\mathscr{D}}[e_k^2] \\
&\leq n^2 - \mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[e_1^2]
\end{aligned}$$

Write e_1 as $n - e_2 - \cdots - e_d$, and we get

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[e_1^2] &= \mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}\left[\left(n - \sum_{i=2}^{d} e_i\right)^2\right] \\ &= n^2 - 2n\sum_{i=2}^{d}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[e_i] + \sum_{i=2}^{d}\sum_{j=2}^{d}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[e_i e_j] \\ &\geq n^2 - 2n\sum_{i=2}^{d}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}}[e_i].\end{aligned}$$

It follows that $\sum_{i\neq j}\mathbb{E}_{\mathbf{e}\sim\mathscr{D}}[e_i e_j] \leq 2n\sum_{i=2}^{d}\mathbb{E}_{\mathbf{e}\sim\mathscr{D}}[e_i]$. Under the approximate distribution, this is

$$\sum_{i=2}^{d}\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}'}[e_i] = \sum_{i=2}^{d}\frac{\lambda_i}{\lambda_1 - \lambda_i} \leq \frac{1-\lambda_1}{2\lambda_1 - 1},$$

and Theorem 36 bounds the difference from the true distribution by at most $\mathcal{O}(\Delta)$. The result follows. ■

B.8. Conclusion

We finish the section by stating and proving a more formal version of Theorem 2.

Theorem 44 *For unknown state with deviation η, the standard joint measurement on n copies succeeds with probability at least $(1-\eta)^{n-1}$. Conditioned on success, there is an estimator $\hat{\phi}$ such that*

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}(O\hat{\phi})] &= \mathrm{Tr}(O\Phi_1) + \frac{1}{n}\cdot\left(\frac{\mathrm{Tr}(O\rho) - \mathrm{Tr}(O\Phi_1)}{1-\eta} + \mathcal{O}\big(\|O\|_\infty \eta^2\big)\right) \\ \mathrm{Var}[\mathrm{Tr}(O\hat{\phi})] &= \frac{\mathrm{Tr}(O^2)}{n^2} + \frac{6\|O\|_\infty^2}{n} + \frac{8\|O\|_\infty^2}{n^2} + \mathcal{O}(\Delta)\end{aligned}$$

Proof Take $\hat{\phi} = \hat{M}_1$. Theorem 38 gives the mean under the approximate distribution

$$\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}'}[M_1^{\mathbf{e}}] = \Phi_1 + \frac{1}{n}\sum_{j=2}^{d}\tfrac{\lambda_j}{\lambda_1-\lambda_j}(\Phi_j - \Phi_1).$$

First, expand $\frac{\lambda_j}{\lambda_1-\lambda_j}$ as:

$$\frac{\lambda_j}{\lambda_1-\lambda_j} = \frac{\lambda_j}{\lambda_1}\left(1 + \mathcal{O}\left(\frac{\lambda_j}{\lambda_1}\right)\right) = \frac{\lambda_j}{\lambda_1}\left(1 + \mathcal{O}\left(\frac{\eta}{1-\eta}\right)\right) = \frac{\lambda_j}{\lambda_1}(1 + \mathcal{O}(\eta)).$$

The first order term gives

$$\sum_{j=2}^{d}\frac{\lambda_j}{\lambda_1}(\Phi_j - \Phi_1) = \sum_{j=1}^{d}\frac{\lambda_j}{\lambda_1}(\Phi_j - \Phi_1) = \rho - \Phi_1.$$

The second order term is the same, but multiplied by $\mathcal{O}(\eta)$, and since

$$\mathrm{Tr}(O(\rho - \Phi_1)) \leq \|O\|_\infty \|\rho - \Phi_1\|_{\mathrm{tr}} = \mathcal{O}(\|O\|_\infty \eta),$$

the expectation is indeed

$$\mathop{\mathbb{E}}_{\mathbf{e}\sim\mathscr{D}'}[\mathrm{Tr}(OM_1^{\mathbf{e}})] = \mathrm{Tr}(O\Phi_1) + \frac{1}{n}\left(\frac{\mathrm{Tr}(O\rho) - \mathrm{Tr}(O\Phi_1)}{1-\eta} + \mathcal{O}\big(\|O\|_\infty \eta^2\big)\right).$$

This is the expectation under $\mathscr{D}'$, but Corollary 39 proves the approximation changes the trace distance by at most $\mathcal{O}(\Delta) \subseteq \mathcal{O}\big(\eta^2\big)$, and thus affects the final expectation by $\mathcal{O}\big(\|O\|_\infty \eta^2\big)$.

On the variance side, Theorem 40 divides the variance into a sum of mixture randomness and measurement randomness. Theorem 41 bounds the mixture randomness:

$$\mathop{\mathrm{Var}}_{\mathbf{e}\in\mathscr{D}}[\mathop{\mathbb{E}}_{\mathrm{meas}}(\mathrm{Tr}(O\hat{M}_1) \mid \mathbf{e})] \leq \frac{4\|O\|^2}{n^2}\frac{\lambda_1(1-\lambda_1)}{(2\lambda_1-1)^2} + \mathcal{O}(\Delta).$$

A combination of Lemma 42 and Lemma 43 bounds the measurement randomness:

$$\mathop{\mathrm{Var}}_{\mathrm{meas}}[\mathrm{Tr}(O\hat{M}_1) \mid \mathbf{e}] \leq \frac{\mathrm{Tr}(O^2)}{n^2} + \frac{2\|O\|_\infty^2}{n} + \frac{2}{n}\frac{1-\lambda_1}{2\lambda_1-1}\|O\|_\infty^2 + \mathcal{O}(\Delta).$$

The total is

$$\begin{aligned}\mathrm{Var}[\mathrm{Tr}(O\hat{\rho})^2] &\leq \frac{\mathrm{Tr}(O^2)}{n^2} + \frac{2\|O\|_\infty^2}{n} + \frac{2\|O\|_\infty^2}{n}\frac{1-\lambda_1}{2\lambda_1-1} + \frac{4\|O\|_\infty^2}{n^2}\frac{\lambda_1(1-\lambda_1)}{(2\lambda_1-1)^2} + \mathcal{O}(\Delta) \\ &\leq \frac{\mathrm{Tr}(O^2)}{n^2} + \frac{6\|O\|_\infty^2}{n} + \frac{8\|O\|_\infty^2}{n^2} + \mathcal{O}(\Delta).\end{aligned}$$

■

Appendix C. Chiribella's Theorem

In this section we adapt Chiribella's theorem Chiribella (2011) to get expressions for $\mathbb{E}[\Psi]$ and $\mathrm{Var}[\Psi]$. The subject of this theorem is the map $\mathrm{MP}_{n\to k}$ defined below.

Definition 45 *For integers $n, k \geq 0$, let $\mathrm{MP}_{n\to k}\colon \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big) \to \mathcal{L}\big((\mathbb{C}^d)^{\otimes k}\big)$ be such that*

$$\mathop{\mathrm{MP}}_{n\to k}(A) = \frac{d_n}{d_{n+k}}\mathop{\mathrm{Tr}}_{[n]}(\Pi_{\mathrm{sym}}^{(n+k)}(A \otimes \mathbb{I}^{\otimes k})) = \frac{d_n}{d_{n+k}}\mathop{\mathrm{Tr}}_{n+k\to k}(A \otimes I^{\otimes k})$$

for all $A \in \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big)$. We remind the reader that $\mathrm{Tr}_{[n]}$ is the partial trace over qudits $[n] = \{1, \ldots, n\}$.

This map is an example of a "measure and prepare map" because it is equivalent to measuring the state with some POVM, and then preparing a state dependent on the outcome. In particular, Proposition 46 below shows that this map measures with $\mathcal{M}_n$, and prepares $|\psi\rangle\langle\psi|^{\otimes k}$ if the outcome is ψ, or 0 if the measurement fails.

Proposition 46 *Let Ψ be the outcome of measuring an n-qudit state A with $\mathcal{M}_s$ ($|\psi\rangle\langle\psi|$ or 0 for failure). Then the k^{th} moment of Ψ is $\mathbb{E}[\Psi^{\otimes k}] = \mathrm{MP}_{n\to k}(A)$ for all $k \geq 0$.*

Proof The expectation of $\Psi^{\otimes k}$ is a straightforward calculation using definition of $\mathcal{M}_n$ (Definition 19) and the Haar integral characterization of $\Pi_{\mathrm{sym}}^{(n)}$ (Lemma 18) as needed.

$$\begin{aligned}
\mathbb{E}[\Psi^{\otimes k}] &= 0^{\otimes k}\operatorname{Tr}(F_\perp A) + \int |\psi\rangle\langle\psi|^{\otimes k}\operatorname{Tr}(F_\psi A) \\
&= \int |\psi\rangle\langle\psi|^{\otimes k} d_n \operatorname{Tr}(|\psi\rangle\langle\psi|^{\otimes n} A)\mathrm{d}\psi && \text{definition of } F_\psi \\
&= d_n \operatorname*{Tr}_{[n]}\left(\left(\int |\psi\rangle\langle\psi|^{\otimes(n+k)}\mathrm{d}\psi\right)(A\otimes\mathbb{I}^{\otimes k})\right) && \text{linearity of trace, integral} \\
&= \frac{d_n}{d_{n+k}}\operatorname*{Tr}_{[n]}\left(\Pi_{\mathrm{sym}}^{(n+k)}(A\otimes\mathbb{I}^{\otimes k})\right) \\
&= \operatorname*{MP}_{n\to k}(A). && \text{definition of MP}
\end{aligned}$$

■

In addition to $\mathrm{MP}_{n\to k}$, Chiribella's theorem uses a "cloning map", defined below.

Definition 47 (Optimal Cloning Map Werner (1998)) *Let us define the superoperator $\mathrm{Cl}_{n\to n+k}\colon \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big) \to \mathcal{L}\big((\mathbb{C}^d)^{\otimes n+k}\big)$ on input $A \in \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big)$ as*

$$\operatorname*{Cl}_{n\to n+k}(A) = \frac{d_n}{d_{n+k}}\Pi_{\mathrm{sym}}^{(n+k)}\Big(A\otimes\mathbb{I}^{\otimes k}\Big)\Pi_{\mathrm{sym}}^{(n+k)}.$$

This map extends an n-qubit state to $n+k$ qudits. The no-cloning theorem prohibits cloning quantum states, but Werner Werner (1998) showed that it is the optimal with respect to the fidelity of $\mathrm{Cl}_{n\to n+k}(\sigma^{\otimes n})$ and $\sigma^{\otimes n+k}$.

This brings us to the key result of this section, due to Chiribella (2011).

Theorem 48 (Chiribella's theorem) *For $A \in \Pi_{\mathrm{sym}}^{(n)}((\mathbb{C}^d)^{\otimes n})$ (in the symmetric subspace)*

$$\operatorname*{MP}_{n\to k}(A) = \binom{d+k+n-1}{k}^{-1}\sum_{s=0}^{k}\binom{n}{s}\binom{d+k-1}{k-s}\operatorname*{Cl}_{s\to k}\left(\operatorname*{Tr}_{[n-s]}(A)\right).$$

Before we get into the relevance of this theorem, let us quickly upgrade it from symmetric states to exchangeable states.

Corollary 49 *For exchangeable $A \in \mathcal{L}\big((\mathbb{C}^d)^{\otimes n}\big)$,*

$$\operatorname*{MP}_{n\to k}(A) = \binom{d+k+n-1}{k}^{-1}\sum_{s=0}^{k}\binom{n}{s}\binom{d+k-1}{k-s}\operatorname*{Cl}_{s\to k}\Big(\operatorname*{Tr}_{n\to s}(A)\Big),$$

where $\mathrm{Tr}_{n\to s}(A) = \mathrm{Tr}_{[n-s]}(\Pi_{\mathrm{sym}}^{(n)}A)$ is taken from Section A.

Proof Set $A' = \Pi_{\mathrm{sym}}^{(n)} A$ and observe that A' is symmetric, since $\Pi_{\mathrm{sym}}^{(n)}$ clearly absorbs permutations on the left, and

$$A' W_\pi = \Pi_{\mathrm{sym}}^{(n)} A W_\pi = \Pi_{\mathrm{sym}}^{(n)} W_\pi^\dagger A = \Pi_{\mathrm{sym}}^{(n)} A = A'$$

on the left, using the exchangeability of A. It follows that we can plug A' into Theorem 48, to get that the right hand side of the claim equals $\mathrm{MP}_{n\to k}(A')$. Then we expand with Definition 45 and see that the extra $\Pi_{\mathrm{sym}}^{(n)}$ can be absorbed into $\Pi_{\mathrm{sym}}^{(n+k)}$:

$$\underset{n\to k}{\mathrm{MP}}(A') = \frac{d_n}{d_{n+k}} \underset{[n]}{\mathrm{Tr}}(\Pi_{\mathrm{sym}}^{(n+k)}(\Pi_{\mathrm{sym}}^{(n)} A \otimes \mathbb{I}^{\otimes k})) = \frac{d_n}{d_{n+k}} \underset{[n]}{\mathrm{Tr}}(\Pi_{\mathrm{sym}}^{(n+k)}(A \otimes \mathbb{I}^{\otimes k})) = \underset{n\to k}{\mathrm{MP}}(A).$$

The result follows. ∎

The relevance of Chiribella's theorem (or the corollary) is that it expresses $\mathrm{MP}_{n\to k}(A)$, and thus $\mathbb{E}[\Psi^{\otimes k}]$ (by Proposition 46), in terms of a handful of partial traces. It distills the state A down to $\leq k$ qudits by partial trace, then blows it back up to k qudits with the cloning map. In other words, we can compute $\mathbb{E}[\Psi]$ and $\mathbb{E}[\Psi^{\otimes 2}]$ entirely from 1-qudit and 2-qudit summaries (i.e., $\mathrm{Tr}_{[n-1]}(\Pi_{\mathrm{sym}}^{(n)} A)$ and $\mathrm{Tr}_{[n-2]}(\Pi_{\mathrm{sym}}^{(n)} A)$) of the full n-qudit state A.

We refactor Chiribella one more time to (i) explicitly link the calculation to $\mathbb{E}[\Psi^{\otimes k}]$, (ii) simplify the binomial coefficients as much as possible, and (iii) expand $\mathrm{Cl}_{k\to s}$ with its definition so that it is not needed in the main text. This is the version of Chiribella's theorem we quote in Section B.1.

Theorem 21 *Fix integers $n, k \geq 0$, let $A \in \mathrm{Dens}\big((\mathbb{C}^d)^{\otimes n}\big)$ be an exchangeable n-qudit state, and let Ψ be the outcome of measuring A with $\mathcal{M}_n$.*

$$\mathbb{E}[\Psi^{\otimes k} \mid \text{success}] = \frac{1}{(d+n)_k} \Pi_{\mathrm{sym}}^{(k)} \left(\sum_{s=0}^{k} \binom{n}{s} \binom{k}{s} \left(\frac{\mathrm{Tr}_{n\to s}(A)}{\mathrm{Tr}_{n\to 0}(A)} \otimes \mathbb{I}^{\otimes k-s} \right) \right) \Pi_{\mathrm{sym}}^{(k)} \tag{8}$$

Proof Start from the definition of MP, plug in the definition of Cl, and expand the binomials to simplify.

$$\begin{aligned}
\underset{n\to k}{\mathrm{MP}}(A) &= \binom{d+k+n-1}{k}^{-1} \sum_{s=0}^{k} \binom{n}{s} \binom{d+k-1}{k-s} \underset{s\to k}{\mathrm{Cl}}\big(\underset{n\to s}{\mathrm{Tr}}(A)\big) \\
&= \binom{d+k+n-1}{k}^{-1} \Pi_{\mathrm{sym}}^{(k)} \left(\sum_{s=0}^{k} \binom{n}{s} \binom{d+k-1}{k-s} \frac{d_s}{d_k} \left(\underset{n\to s}{\mathrm{Tr}}(A) \otimes \mathbb{I}^{\otimes k-s} \right) \right) \Pi_{\mathrm{sym}}^{(k)} \\
&= \frac{1}{(d+n)_k} \Pi_{\mathrm{sym}}^{(k)} \left(\sum_{s=0}^{k} \binom{n}{s} \frac{(d+k-1)!}{(k-s)!(d+s-1)!} \frac{\frac{(d+s-1)!}{(d-1)!s!}}{\frac{(d+k-1)!}{(d-1)!k!}} \left(\underset{n\to s}{\mathrm{Tr}}(A) \otimes \mathbb{I}^{\otimes k-s} \right) \right) \Pi_{\mathrm{sym}}^{(k)} \\
&= \frac{1}{(d+n)_k} \Pi_{\mathrm{sym}}^{(k)} \left(\sum_{s=0}^{k} \binom{n}{s} \binom{k}{s} \left(\underset{n\to s}{\mathrm{Tr}}(A) \otimes \mathbb{I}^{\otimes k-s} \right) \right) \Pi_{\mathrm{sym}}^{(k)}.
\end{aligned}$$

Recall that $\mathbb{E}[\Psi^{\otimes k}] = \mathrm{MP}_{n\to k}(A)$ for exchangeable A by Theorem 49. Since

$$\begin{aligned}\mathbb{E}[\Psi^{\otimes k}] &= \mathbb{E}[\Psi^{\otimes k} \mid \text{success}]\Pr[\text{success}] + \mathbb{E}[\Psi^{\otimes k} \mid \neg\text{success}]\Pr[\neg\text{success}] \\ &= \mathbb{E}[\Psi^{\otimes k} \mid \text{success}]\operatorname{Tr}(\Pi_{\mathrm{sym}}^{(n)} A),\end{aligned}$$

we can divide through by $\mathrm{Tr}_{n\to 0}(A) := \operatorname{Tr}(\Pi_{\mathrm{sym}}^{(n)} A)$ to get the expectation of $\Psi^{\otimes k}$ conditioned on success. ■

Appendix D. Approximating the distribution of e

This appendix is dedicated to results about the approximate distribution $\mathscr{D}'$, and how it relates to $\mathscr{D}$ and the expectation values we wish to compute .

First, we recall the mean and variance of the geometric random variables composing $\mathbf{e} \sim \mathscr{D}'$.

Proposition 50 *Let $\mathbf{e} \sim \mathscr{D}'$. The mean and variance of e_i for $2 \le i \le d$ is*

$$\mathbb{E}[e_i] = \frac{\lambda_i}{\lambda_1 - \lambda_i}, \qquad \mathrm{Var}[e_i] = \frac{\lambda_i \lambda_1}{(\lambda_1 - \lambda_i)^2},$$

Next, we want to bound the difference between $\mathscr{D}$ and $\mathscr{D}'$. To start, we bound the probability $e_i = n - j$ for arbitrary j.

Lemma 51 *When $\mathbf{e} \sim \mathscr{D}'$, we have* $\Pr[e_2 + \cdots + e_d = j = n - e_1] \le \left(\frac{1-\lambda_1}{\lambda_1}\right)^j$.

Proof From the definitions, we have

$$\Pr[e_2 + \cdots + e_d = j] = \sum_{e_2+\cdots+e_d=j} \prod_{i=2}^{d} f_i(e_i) = \sum_{e_2+\cdots+e_d=j} \prod_{i=2}^{d} (\tfrac{\lambda_i}{\lambda_1})^{e_i}(1 - \tfrac{\lambda_i}{\lambda_1}).$$

The factors $(1 - \frac{\lambda_i}{\lambda_1})$ are all ≤ 1 and can be neglected. Then we introduce multinomial coefficients $\binom{j}{\mathbf{e}} \ge 1$ into the sum, letting us apply the multinomial theorem.

$$\Pr[e_2 + \cdots + e_d = j] \le \sum_{e_2+\cdots+e_d=j} \binom{j}{\mathbf{e}} \prod_{i=2}^{d} (\tfrac{\lambda_i}{\lambda_1})^{e_i} = \left(\sum_{i=2}^{d} \frac{\lambda_i}{\lambda_1}\right)^j = \left(\frac{1-\lambda_1}{\lambda_1}\right)^j.$$

■

This leads to a bound on the probability e_1 is negative, which then also bounds the distance between the two distributions.

Theorem 35

$$\Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 < 0] \le \Delta := \left(\frac{1-\lambda_1}{\lambda_1}\right)^{n+1} \frac{\lambda_1}{2\lambda_1 - 1}.$$

It follows that $\|\mathscr{D} - \mathscr{D}'\|_{TV} = \Delta$.

Proof Recall that $e_1 = n - e_2 - \cdots - e_d$, so

$$\Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 < 0] = \Pr_{\mathbf{e}\sim\mathscr{D}'}[e_2 + \cdots + e_d > n] = \sum_{j=n+1}^{\infty} \Pr_{\mathbf{e}\sim\mathscr{D}'}[e_2 + \cdots + e_d = j].$$

Use Lemma 51 and sum a geometric series to get

$$\Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 < 0] \leq \sum_{j=n+1}^{\infty} \left(\frac{1-\lambda_1}{\lambda_1}\right)^j = \left(\frac{1-\lambda_1}{\lambda_1}\right)^{n+1} \frac{\lambda_1}{2\lambda_1 - 1} = \Delta.$$

We know that where $\mathscr{D}$ has support, the mass is proportional to $\mathscr{D}'$, but necessarily larger because $\mathscr{D}'$ has mass on $e_1 < 0$ where $\mathscr{D}$ does not. Moreover, $e_1 < 0$ is the only area where $\mathscr{D}'$ has support and $\mathscr{D}$ does not. Thus, $\Pr[e_1 < 0]$ is exactly the mass which must be moved to transform $\mathscr{D}'$ into $\mathscr{D}$, and hence $\|\mathscr{D} - \mathscr{D}'\|_{TV} = \Delta$. ■

We discuss in the main text that $\Delta = \mathcal{O}(\eta^2)$, and can be *very* small for practical values of λ_1 and n.

Recall that we need $\mathbb{E}_{\mathbf{e}\sim\mathscr{D}}[e_i]$ and $\mathbb{E}_{\mathbf{e}\sim\mathscr{D}}[e_i e_j]$ to evaluate M_1 and M_2. The total variation distance alone is insufficient to bound the difference in expectation, so we must separately justify how much the approximation can distort expectations.

Theorem 36 *The difference between $\mathscr{D}$ and $\mathscr{D}'$ for first-order expectations is at most*

$$\left\| \mathop{\mathbb{E}}_{\mathscr{D}}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] \right\|_1 = \sum_i \left| \mathop{\mathbb{E}}_{\mathscr{D}}[e_i] - \mathop{\mathbb{E}}_{\mathscr{D}'}[e_i] \right| \leq \frac{2\Delta}{1-\Delta}\left(n + \frac{1}{2\lambda_1 - 1}\right).$$

Proof For some small probability $p := \Pr_{\mathscr{D}'}[e_1 < 0] \leq \Delta$, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] &= \Pr[e_1 \geq 0] \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e} \mid e_1 \geq 0] + \Pr[e_1 < 0] \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e} \mid e_1 < 0] \\ &= (1-p) \mathop{\mathbb{E}}_{\mathscr{D}}[\mathbf{e}] + p \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e} \mid e_1 < 0]. \end{aligned}$$

Rearranging, $\mathbb{E}_{\mathscr{D}}[\mathbf{e}] = \frac{1}{1-p}(\mathbb{E}_{\mathscr{D}'}[\mathbf{e}] - p\,\mathbb{E}_{\mathscr{D}'}[\mathbf{e} \mid e_1 < 0])$, and thus

$$\left\| \mathop{\mathbb{E}}_{\mathscr{D}}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] \right\|_1 = \frac{p}{1-p} \left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e} \mid e_1 < 0] \right\|_1.$$

Since the sum of $\mathbf{e}$ is always n under any of these distributions, the expectations of $\mathbf{e}$ also sum to n. The sum of the coordinate-wise differences is $n - n = 0$, and thus the absolute difference on e_1 is, by triangle inequality, bounded by the absolute differences on for the other coordinates. That is,

$$\left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e} \mid e_1 < 0] \right\|_1 \leq 2 \left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1} \mid e_1 < 0] \right\|_1 \leq 2 \left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1}] \right\|_1 + 2 \left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1} \mid e_1 < 0] \right\|_1,$$

where $\mathbf{e}_{-1} = (e_2, \ldots, e_d)$.

Since $e_2, \dots, e_d \geq 0$, these norms are both just the sum of the entries, i.e.,

$$\left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1}] \right\|_1 = \sum_{i=2}^{d} \mathop{\mathbb{E}}_{\mathscr{D}'}[e_i].$$

For these geometric random variables, we have $\mathbb{E}[e_i] = \frac{\lambda_i}{\lambda_1 - \lambda_i} \leq \frac{\lambda_i}{2\lambda_1 - 1}$, for all $2 \leq i \leq d$. This gives the following bound on the norm.

$$\left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1}] \right\|_1 = \sum_{i=2}^{d} \mathbb{E}[e_i] \leq \frac{\sum_{i=2}^{d} \lambda_i}{2\lambda_1 - 1} = \frac{1 - \lambda_1}{2\lambda_1 - 1}.$$

On the other hand,

$$\begin{aligned}
\Pr[e_1 < 0] \mathop{\mathbb{E}}_{\mathbf{e} \sim \mathscr{D}'}[e_2 + \cdots + e_d \mid e_1 < 0] &= \Pr[e_2 + \cdots + e_d > n] \mathop{\mathbb{E}}_{\mathbf{e} \sim \mathscr{D}'}[e_2 + \cdots + e_d \mid e_2 + \cdots + e_d > n] \\
&= \sum_{j=n+1}^{\infty} j \cdot \Pr[e_2 + \cdots + e_d = j] \\
&\leq \sum_{j=n+1}^{\infty} j \left(\frac{1 - \lambda_1}{\lambda_1} \right)^j \\
&= \left(\frac{1 - \lambda_1}{\lambda_1} \right)^{n+1} \frac{\lambda_1}{(2\lambda_1 - 1)^2} (n(2\lambda_1 - 1) + \lambda_1) \\
&= \Delta \left(n + \frac{\lambda_1}{2\lambda_1 - 1} \right)
\end{aligned}$$

$$\begin{aligned}
\left\| \mathop{\mathbb{E}}_{\mathscr{D}}[\mathbf{e}] - \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}] \right\|_1 &= \frac{2p}{1 - p} \left(\left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1}] \right\|_1 + \left\| \mathop{\mathbb{E}}_{\mathscr{D}'}[\mathbf{e}_{-1} \mid e_1 < 0] \right\|_1 \right) \\
&\leq \frac{2\Delta}{1 - \Delta} \left(\frac{1 - \lambda_1}{2\lambda_1 - 1} + n + \frac{\lambda_1}{2\lambda_1 - 1} \right) \\
&= \frac{2\Delta}{1 - \Delta} \left(n + \frac{1}{2\lambda_1 - 1} \right)
\end{aligned}$$

■

Finally, we bound the variance by going through the covariance matrices.

Lemma 52 *The covariance matrices of $\mathscr{D}$ and $\mathscr{D}'$ are related as follows:*

$$\Sigma_{\mathscr{D}'} \succeq \Sigma_{\mathscr{D}} (1 - \Delta)^2.$$

Proof An elegant way to write the covariance matrix is

$$\Sigma_{\mathscr{D}'} = \mathop{\mathbb{E}}_{\substack{\mathbf{e} \sim \mathscr{D}' \\ \mathbf{e}' \sim \mathscr{D}'}} [(\mathbf{e} - \mathbf{e}')(\mathbf{e} - \mathbf{e}')^\top].$$

Each $(\mathbf{e}-\mathbf{e}')(\mathbf{e}-\mathbf{e}')^\top$ is positive semidefinite, and therefore so is $\Sigma_{\mathscr{D}'}$. Now split the expectation based on whether e_1 and e_1' are nonnegative.

$$\Sigma_{\mathscr{D}'} = \mathop{\mathbb{E}}_{\substack{\mathbf{e}\sim\mathscr{D}'\\ \mathbf{e}'\sim\mathscr{D}'}}[(\mathbf{e}-\mathbf{e}')(\mathbf{e}-\mathbf{e}')^\top \mid e_1 \geq 0 \wedge e_1' \geq 0] \Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 \geq 0]^2 + \mathop{\mathbb{E}}_{\substack{\mathbf{e}\sim\mathscr{D}'\\ \mathbf{e}'\sim\mathscr{D}'}}[(\mathbf{e}-\mathbf{e}')(\mathbf{e}-\mathbf{e}')^\top \mid e_1 < 0 \vee e_1' < 0]\left(1 - \Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 \geq 0]^2\right)$$

Recall that $\mathscr{D}$ is $\mathscr{D}'$ conditioned on $e_1 \geq 0$ (Lemma 34), so that first conditional expectation is actually $\Sigma_{\mathscr{D}}$.

$$\mathop{\mathbb{E}}_{\substack{\mathbf{e}\sim\mathscr{D}'\\ \mathbf{e}'\sim\mathscr{D}'}}[(\mathbf{e}-\mathbf{e}')(\mathbf{e}-\mathbf{e}')^\top \mid e_1 \geq 0 \wedge e_1' \geq 0] = \mathop{\mathbb{E}}_{\substack{\mathbf{e}\sim\mathscr{D}\\ \mathbf{e}'\sim\mathscr{D}}}[(\mathbf{e}-\mathbf{e}')(\mathbf{e}-\mathbf{e}')^\top] = \Sigma_{\mathscr{D}}$$

It follows that

$$\Sigma_{\mathscr{D}'} \succeq \Sigma_{\mathscr{D}} \Pr_{\mathbf{e}\sim\mathscr{D}'}[e_1 \geq 0]^2 \succeq \Sigma_{\mathscr{D}}(1-\Delta)^2.$$

■

Appendix E. Optimal Parameter Choice

We now prove Theorem 9.

Theorem 9 *Given B, ϵ and η, the expected number of samples is minimized for the choice of k, n and b given in Table 1.*

For convenience, we duplicate Table 1 below.

η	$\mathcal{O}(1/s^*)$	$\Omega(1/s^*) \cap \mathcal{O}(\sqrt{\epsilon})$	$\Omega(\sqrt{\epsilon})$
k	1	1	$\mathcal{O}\left(\frac{\eta}{\sqrt{\epsilon}}\right)$
n	$\mathcal{O}(s^*)$	$\mathcal{O}\left(\frac{1}{\eta}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$
b	1	$\mathcal{O}\left(\frac{B\eta^2+\eta}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{B}{\epsilon}+\frac{1}{\epsilon^{3/2}}\right)$
s	$\mathcal{O}(s^*)$	$\mathcal{O}\left(\frac{B\eta+1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{B\eta}{\epsilon^2}+\frac{\eta}{\epsilon^{5/2}}\right)$

Table 2: Choice of parameters k, n, b for the three regimes of η. Recall $s^* := \frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}$.

Proof

The optimal choice of parameters k, n, b is essentially a mathematical program. The objective is s, the expected number of samples of ρ. Figure 1 makes it clear that k, n, b multiply, and since that only counts the *successful* measurements, we furthermore multiply by a factor $1/\Pr[\text{success}]$ in expectation. We recall that $\Pr[\text{success}]$ falls off exponentially with n (cf. Equation (14)). It is therefore advisable to keep $\Pr[\text{success}]$ above a constant, which is achieved by setting $n = \mathcal{O}(1/\eta') = \mathcal{O}(k/\eta)$. This gives us both our objective value, $s = knb/\Pr[\text{success}] = \Theta(knb)$, and first constraint, Equation (19).

The remaining constraints come from correctness, i.e., the requirement to output an estimator with error at most ϵ. The mean squared error of our estimate is split between the bias squared and the variance. From Equation (3) and Equation (5), we deduce that the bias β is at most $\mathcal{O}\big(\frac{\eta}{kn}\big)$. From Theorem 2, the total variance of our estimator satisfies: $\mathrm{Var}(\mathrm{Tr}(O\hat{\phi}(b))) = \mathcal{O}(\frac{B}{n^2 b} + \frac{1}{nb})$. To ensure the mean squared error is at most $\mathcal{O}\big(\epsilon^2\big)$, the bias should be $\mathcal{O}(\epsilon)$, and each term of the variance at most $\mathcal{O}\big(\epsilon^2\big)$ (or standard deviation at most $\mathcal{O}(\epsilon)$). The variance terms translate directly to constraints (17), and (18), and the bias condition gives (16). Last, we require k, n, b to be at least 1.

$$\begin{aligned}
\text{minimize} \quad & knb \\
\text{subject to} \quad & \frac{\eta}{nk} = \mathcal{O}(\epsilon) && \text{(bias condition)} && (16)\\
& \frac{B}{n^2 b} = \mathcal{O}\big(\epsilon^2\big) && \text{(variance condition 1)} && (17)\\
& \frac{1}{nb} = \mathcal{O}\big(\epsilon^2\big) && \text{(variance condition 2)} && (18)\\
& n = \mathcal{O}(k/\eta) && \text{(success condition)} && (19)\\
& k, n, b \geq 1. && \text{(positivity condition)} && (20)
\end{aligned}$$

It remains to optimize this program for arbitrary B, ϵ, η.

Table 1 gives optimal values for k, n, and b in each of the three regimes. It is a calculation to see that these solutions are feasible and achieve the claimed sample complexities. On the other hand, optimality can be certified by the following products of constraints.

$$\begin{aligned}
\sqrt{(17)} \times (1 \leq \sqrt{b}) \times (1 \leq k) &\implies knb = \Omega(\sqrt{B}/\epsilon) && (21)\\
(18) \times (1 \leq k) &\implies knb = \Omega(1/\epsilon^2) && (22)\\
(17) \times (19) &\implies knb = \Omega(B\eta/\epsilon^2) && (23)\\
\sqrt{(16)} \times (18) \times \sqrt{(19)} &\implies knb = \Omega(\eta/\epsilon^{5/2}) && (24)
\end{aligned}$$

For example, if we multiply $1/nb = \mathcal{O}(\epsilon^2)$ and $1 \leq k$ we get $1/nb = \mathcal{O}(k\epsilon^2)$, which we can rearrange to $knb = \Omega(1/\epsilon^2)$, (22). It is clear from these equations that the complexities of the three regimes—$\mathcal{O}(\sqrt{B}/\epsilon + 1/\epsilon^2)$, $\mathcal{O}((B\eta + 1)/\epsilon^2)$, and $\mathcal{O}(B\eta/\epsilon^2 + \eta/\epsilon^{5/2})$—arise from (21) + (22), (22) + (23), and (23) + (24) respectively. Likewise, thresholds between regimes are given by the crossover points of these inequalities, i.e., $\sqrt{B}/\epsilon \approx B\eta/\epsilon^2$ implies $\eta \approx \epsilon/\sqrt{B} = \Theta(1/s^*)$ and $1/\epsilon^2 \approx \eta/\epsilon^{5/2}$ gives $\eta \approx \sqrt{\epsilon}$. ■

Appendix F. Eigenvalue estimation

In this appendix, we prove the following theorem, as a component of estimating η from samples of ρ.

Theorem 10 *Let $r \geq 1$ be an integer. There is an algorithm which estimates the failure probability p of a Bernoulli trial, such that the algorithm (i) outputs a constant-factor multiplicative approximation of p, and (ii) makes $\mathcal{O}(r/p)$ samples of the Bernoulli trial, except with an $\exp(-\Theta(r))$ probability of failure.*

The algorithm is simple: it performs/samples Bernoulli trials until it has seen r failures total. Then it outputs r/T as an estimate for p, where T is the total number of *trials*.

Lemma 53 *The number of trials T is bounded above and below with high probability, i.e.,*

$$\Pr[T = \Theta(r/p)] \geq 1 - \exp(-\Theta(r)).$$

Concretely, we have, e.g.,

$$\Pr[\ln(2)\tfrac{r}{p} \leq T \leq \ln(4)\tfrac{r}{p}] \geq 1 - 2 \cdot (\tfrac{1}{2}e \ln 2)^r \geq 1 - 2 \cdot 0.9421^r.$$

Proof First, we argue that for any $n \geq r$, if $Y \sim \mathrm{Bin}(n, p)$ then $\Pr[T \leq n] = \Pr[Y \geq r]$. To see this, imagine an infinite sequence of Bernoulli trials. The binomial distribution counts the number of failures, Y, in the first n trials, whereas the algorithm scans down the list to the rth failure at some position T. It is clear that the rth failure happens at or before position n ($T \leq n$) if and only if there are r or more failures among the first n trials ($Y \geq r$).

Since the mean of the binomial is $\mathbb{E}[Y] = np$, a multiplicative Chernoff bound gives

$$\Pr[Y \geq (1+x)np] \leq \left(\frac{e^x}{(1+x)^{1+x}}\right)^{np}.$$

Set $n = \frac{r}{(1+x)p}$ so that $r = (1+x)np$, and this becomes

$$\Pr\left[T \leq \frac{r}{(1+x)p}\right] = \Pr[Y \geq r] \leq \left(\frac{e^{x/(1+x)}}{1+x}\right)^r.$$

For instance, at $x^* = \frac{1}{\ln 2} - 1$, we have $\Pr[T \leq \frac{r}{p}\ln 2] \leq (\frac{1}{2}e \ln 2)^r \approx 0.9421^r$.

On the other side, $\Pr[T > n] = \Pr[Y < r] \leq \Pr[Y \leq r]$. The other side of the Chernoff bound gives

$$\Pr[Y \leq (1-x)np] \leq \left(\frac{e^{-x}}{(1-x)^{(1-x)}}\right)^{np}.$$

Setting $r = (1-x)np$, we translate this to

$$\Pr\left[T \geq \frac{r}{(1-x)p}\right] \leq \left(\frac{e^{-x/(1-x)}}{1-x}\right)^r.$$

At $x^* = 1 - \frac{1}{\ln 4}$ we get $\Pr[T \geq \frac{r}{p}\ln 4] \leq (\frac{1}{2}e \ln 2)^r \approx 0.9421^r$. Union bound over the two tail bounds finishes the result. ■