Sample-optimal classical shadows for pure states

Daniel Grier[1,2], Hakop Pashayan[2,3,4,5], and Luke Schaeffer[2,3,6]

[1]Department of Mathematics and Department of Computer Science and Engineering, UC San Diego
[2]Institute for Quantum Computing, University of Waterloo, Canada
[3]Department of Combinatorics and Optimization, University of Waterloo, Canada
[4]Perimeter Institute for Theoretical Physics, Waterloo, Canada
[5]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Germany
[6]Joint Center for Quantum Information and Computer Science, University of Maryland, College Park

We consider the classical shadows task for pure states in the setting of both joint and independent measurements. The task is to measure few copies of an unknown pure state ρ in order to learn a classical description which suffices to later estimate expectation values of observables. Specifically, the goal is to approximate $\mathrm{Tr}(O\rho)$ for any Hermitian observable O to within additive error ϵ provided $\mathrm{Tr}(O^2) \leq B$ and $\|O\| = 1$. Our main result applies to the joint measurement setting, where we show $\tilde{\Theta}(\sqrt{B}\epsilon^{-1} + \epsilon^{-2})$ samples of ρ are necessary and sufficient to succeed with high probability. The upper bound is a quadratic improvement on the previous best sample complexity known for this problem. For the lower bound, we see that the bottleneck is not how fast we can learn the state but rather how much any classical description of ρ can be compressed for observable estimation. In the independent measurement setting, we show that $\mathcal{O}(\sqrt{Bd}\epsilon^{-1} + \epsilon^{-2})$ samples suffice. Notably, this implies that the random Clifford measurements algorithm of Huang, Kueng, and Preskill, which is sample-optimal for mixed states, is not optimal for pure states. Interestingly, our result also uses the same random Clifford measurements but employs a different estimator.

1 Introduction

How many copies of an unknown state are required to construct a classical description of the state? The answer to this question will depend on several details: what constitutes an accurate description; what is already known about the state; and what restrictions are placed on the measurements of the state. Given the fundamental importance of this question, there has been significant prior work in bounding the number of samples of the states required to perform this learning task in a variety of contexts.

The most well-known setting is called *quantum state tomography*, where the goal is to learn enough about the state to be able to completely reconstruct it—precisely, estimate the unknown d-dimensional quantum state to accuracy ϵ in the Schatten 1-norm. Tight upper and lower bounds for the number of copies required for this task are known: $\tilde{\Theta}(\epsilon^{-2}d^3)$ copies of the state are needed with independent measurements [1], and $\tilde{\Theta}(\epsilon^{-2}d^2)$ copies are needed when the

unknown states can be simultaneously measured in a large joint measurement [2]. Independent measurements are easier to experimentally implement, while the joint measurements explore what is possible with respect to the fundamental limits of quantum mechanics. A key takeaway in the joint measurement setting is that the algorithm for the upper bound is achieving what would naïvely be the best possible result, given that a d-dimensional state has $\Theta(d^2)$-many independent parameters, and $\Theta(\epsilon^{-2})$ samples are necessary to estimate any one parameter.

In some sense, the requirements of the quantum tomography question are quite rigid. For many applications, only some properties of the unknown state are important. Can we get away with fewer samples if we relax our notion of approximation? In particular, what if we only wish to learn the expected values of certain Hermitian observables? Aaronson gave a somewhat surprising answer to this question in a joint measurement setting called *shadow tomography* [3]: given M bounded observables $(\{O_i\}_{i=1}^M, \|O_i\| \leq 1)$, estimate $\mathrm{Tr}(O_i\rho)$ to within ϵ additive error.[1] In this setting, Aaronson showed that only $\tilde{\mathcal{O}}(\epsilon^{-4}\log^4 M \log d)$ samples of the state are needed. Subsequent work by Bădescu and O'Donnell [4] improved this to $\tilde{\mathcal{O}}(\epsilon^{-4}\log^2 M \log d)$, but there are still no matching lower bounds for this setting. That is, we do not know if we are extracting as much information about the unknown state as we can. In the independent measurement setting, $\tilde{\Theta}(\min\{M, d\}/\epsilon^2)$ samples are necessary and sufficient [5].

One subtlety concerning these observable estimation tasks is whether or not the measurements are allowed to depend on the specific observables O_i. In shadow tomography, the measurements *can* depend on the observables, but an increasingly popular setting (inspired by the work of Huang, Kueng, and Preskill [6]) is one in which the observables O_i are unknown at the time of measurement. That is, the measurements must produce a classical description (called the *classical shadow*) from which the observable expected values can later be calculated. In their randomized Clifford measurement scheme, Huang, Kueng, and Preskill consider the independent measurement setting and show that $\Theta(B\epsilon^{-2}\log M)$ copies of the unknown state are both necessary and sufficient provided that $\mathrm{Tr}(O_i^2) \leq B$ for all i (note that $\mathrm{Tr}(O_i^2) \leq d$).

Consider now how the classical shadows setting compares to the quantum state tomography setting with regard to the type of measurements allowed. In the quantum state tomography setting, we know that joint measurements allow us to extract more information from the state, yielding estimates of the unknown state with provably fewer samples than those required with independent measurements. In the classical shadows setting, however, it is not known how the type of measurement affects the number of samples required. Concretely, is it possible to perform the classical shadows task with fewer samples if we switch to joint measurements? We answer this question affirmatively in the setting of *pure* states.

Formally, we show the following: $\mathcal{O}((\sqrt{B}\epsilon^{-1} + \epsilon^{-2})\log M)$ samples of the unknown pure state are sufficient for performing the classical shadows task with constant probability of failure. Compared to [6], this achieves almost a square root reduction in sample complexity.

Remarkably, in analogy with the quantum state tomography setting, our joint measurement procedure is in some sense extracting the maximum amount of information possible. To see this, consider a simple setting in which $B = d$, ϵ is constant, and we only wish to estimate

[1] Aaronson actually stipulates that each observable is positive semi-definite matrix E_i so that $\{E_i, I - E_i\}$ is a 2-outcome POVM. We note that this is equivalent to the task of estimating expectation values of the (bounded) Hermitian observables O_i via the mapping $E_i = (O_i + I)/2$.

a single observable. Our algorithm uses $\mathcal{O}(\sqrt{d})$ samples. However, Gosset and Smolin [7] show that even if you are given the state as an explicit density matrix, you cannot compress your description of the state down to fewer than $\Omega(\sqrt{d})$-many bits of information in order to estimate arbitrary observable expectation values. Notice, however, that to successfully execute the classical shadows task, one would first need to learn such a compressed description through measurement of the unknown state. A priori, the number of measurements required to do this could be much higher than the size of this compressed description. The fact that we find a matching upper bound implies that accessing the relevant information contained in the state is not the significant bottleneck.

We show that a similar phenomena exists for arbitrary parameters B and ϵ. Namely, we refine the Gosset-Smolin lower bound for compression to $\Omega(\sqrt{B}\epsilon^{-1})$-many bits, which ultimately allows us to show that $\tilde{\Omega}(\sqrt{B}\epsilon^{-1} + \epsilon^{-2})$ samples of the state are required for the classical shadows task. Therefore, our joint-measurement algorithm above is sample-optimal (at least for a single observable and up to log factors).

Finally, we address the classical shadows question with pure states and independent measurements. We show that $\mathcal{O}((\frac{\sqrt{Bd}}{\epsilon} + \frac{1}{\epsilon^2}) \log M)$ copies of the state suffice. It's worth noticing that in certain parameter regimes, this upper bound is smaller than $\Theta(B\epsilon^{-2} \log M)$. In other words, our algorithm uses fewer samples than the classical shadows algorithm of Huang, Kueng, and Preskill which was designed for general mixed states. Indeed, their lower bound methods require the underlying state to be mixed.

1.1 The classical shadows task

We consider the classical shadows task introduced by Huang, Kueng, and Preskill [6]: given several copies of an unknown quantum state, produce a classical description of the state that is sufficiently representative to permit the reliable and accurate estimation of expectation values of some number of observables chosen from a broad class.

To formalise the task, let's begin with the class of observables we will use:

Definition 1. For any $B \in (0, d]$, let

$$\mathrm{Obs}(B) := \left\{ O \in \mathbb{C}^{d \times d} \mid O = O^\dagger, \|O\|_\infty = 1, \mathrm{Tr}(O^2) \leq B \right\}.$$

In summary, these observables have been scaled/normalized so that $\|O\|_\infty = 1$ and have a bound of B on their squared Frobenius norm $\mathrm{Tr}(O^2)$. The latter condition is due to the fact that $\mathrm{Tr}(O^2)$ is typically the dominant term in the sample complexity. We could also reasonably upper bound it by the rank of the observable since $\mathrm{Tr}(O^2) \leq \mathrm{rank}\, O \leq d$.

We remark that $\|O\|_2 = \sqrt{\mathrm{Tr}(O^2)}$ and $\|O\|_\infty$ are examples of Schatten p-norms where $p = 2$ and $p = \infty$ respectively, but defined in general as $\|A\|_p := \mathrm{Tr}(|A|^p)^{1/p}$ for $p \in [1, \infty)$. We will also use the Schatten 1-norm. Going forward, we write $\|O\|_1$ for the 1-norm, $\|O\|$ for the infinity norm, and prefer $\mathrm{Tr}(O^2)$ over $\|O\|_2^2$.

Definition 2 (Classical Shadows Task)**.** The Classical Shadows Task consists of two separate phases—a measurement phase and an observable estimation phase—which are completed by two separate (randomized) algorithms, $\mathcal{A}_{\mathrm{meas}}$ and $\mathcal{A}_{\mathrm{est}}$, respectively. In addition to the inputs below, each algorithm also depends on the four parameters s, B, ϵ, and δ:

Measurement: $\mathcal{A}_{\text{meas}} \colon \rho^{\otimes s} \to \{0,1\}^*$
Input: s copies of a state $\rho \in \mathbb{C}^{d \times d}$.
Output: A bit string called the *classical shadow*.

Estimation: $\mathcal{A}_{\text{est}} \colon \text{Obs}(B) \times \{0,1\}^* \to \mathbb{R}$
Input: Observable $O \in \text{Obs}(B)$ and a classical shadow.
Output: Estimate $E \in \mathbb{R}$.

It's worth emphasizing that the input to the measurement algorithm is quantum (the state $\rho^{\otimes s}$) and the output is classical (the classical shadow). This output is computed from measuring the input state with some POVM (with arbitrary post-processing).

We say that $\mathcal{A}_{\text{meas}}$ and $\mathcal{A}_{\text{est}}$ constitute a valid protocol for the classical shadows task if their estimate for the expectation of the observable $E := \mathcal{A}_{\text{est}}(O, \mathcal{A}_{\text{meas}}(\rho^{\otimes s}))$ is such that

$$|\text{Tr}(O\rho) - E| < \epsilon \tag{1}$$

with probability at least $1 - \delta$ over the randomness of $\mathcal{A}_{\text{meas}}$ and $\mathcal{A}_{\text{est}}$.

Some may find it useful to think about the classical shadows task as a one-way communication protocol where one party (let's call her Melanie) is given copies of an unknown state and another party (say, Esteban) is given an observable. Melanie doesn't know Esteban's observable, and Esteban cannot send hints because we are assuming one way communication from Melanie to Esteban, so there is only one course of action: Melanie must measure her unknown state and send (over a classical channel) a description of the state from which Esteban can estimate the expected value of his given observable.

Throughout this paper, we will focus on the classical shadows task with unknown *pure* states. This motivates the following definitions:

Definition 3 (Sample Complexity of the Classical Shadows Task)**.** Let $\mathsf{Shadows}(B, \epsilon, \delta)$ to be the minimum number of samples s required to successfully carry out the classical shadows task on pure states with the set of observables $\text{Obs}(B)$, to accuracy ϵ, and failure probability at most δ.

Sometimes we will omit δ and write $\mathsf{Shadows}(B, \epsilon)$ to denote the minimum number of samples to achieve these tasks with some constant probability of failure, say, 0.001.

Definition 4 (Classical Shadows with Independent Measurements)**.** Let $\mathsf{I\text{-}Shadows}(B, \epsilon, \delta)$ be the sample complexity for the classical shadows task with pure states when the measurement algorithm can only make *independent* measurements on the input state—that is, the measurement POVM is the tensor product of POVMs on single copies of the state. These POVMs do not have to be identical, but the entire state must be measured at the same time, or in other words, the output from a measurement on one copy of the state cannot influence the measurement on another.

We note that there are many possible variants for the sample complexity of the classical shadows task that we haven't given individual names. Most notably are the settings where the unknown states are mixed states (rather than pure) and/or the measurements are allowed to be adaptive (while still acting on single copies of the state).

1.2 Summary of results

Our main result is to prove matching upper and lower bounds on the sample complexity of performing the classical shadows task with respect to joint measurements and pure states.

Theorem 5. $\mathsf{Shadows}(B,\epsilon) = \tilde{\Theta}\left(\frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}\right)$ *provided* $B \leq \epsilon d$.

Notice that Theorem 5 consists of separate upper and lower bound results (for constant δ). These match up to logarithmic factors in B and ϵ^{-1}, and the technical relationship between B, ϵ, and d is only required for the lower bound. In Section 3, we will prove the upper bound, where we will also show that the dependence on the failure probability δ goes as $\log(1/\delta)$. We note that this dependence on δ implies that there are efficient protocols for the calculation of several observables simultaneously—that is, if the classical shadows task fails with probability at most δ on a single observable, then it fails with probability at most $M\delta$ on one or more out of M observables by the union bound. In Section 4, we will prove the lower bound where only the ϵ^{-2} term will scale with $\log(1/\delta)$.

We also prove an upper bound on the sample complexity of performing the classical shadows task with respect to independent measurements and *pure* states. Our upper bound can be compared to the matching upper and lower bound of Huang, Kueng, and Preskill [6] which applies to independent measurements and *general* states. In certain parameter regimes, our upper bound achieves a *smaller* sample complexity than the lower bound in [6] which implies that in the independent measurement setting, the classical shadows task has smaller sample complexity for pure states.

Theorem 6. *For all* $\epsilon, \delta > 0$,

$$\mathsf{I\text{-}Shadows}(B,\epsilon,\delta) = \mathcal{O}\left(\min\left\{\frac{B}{\epsilon^2}, \frac{\sqrt{Bd}}{\epsilon} + \frac{1}{\epsilon^2}\right\}\log(\delta^{-1})\right).$$

We discuss and prove Theorem 6 in Section 5.

Finally, we note that in all of our algorithms, the estimator $\hat{\rho}$ we use for the unknown state ρ is not itself a proper state. In Appendix A, we show that this is a necessary price for the favorable sample complexity enjoyed by classical shadows schemes. Informally, we show that even for observables in $\mathrm{Obs}(1)$, learning an estimate $\hat{\rho}$ that is a proper state to sufficient accuracy to solve the classical shadows task via the formula $\mathrm{Tr}(O\hat{\rho})$, requires a sample complexity that scales linearly in d, the dimension of the unknown state.

2 Preliminaries

Here we cover key background material related to Haar random states, their moments, and the symmetric subspace. Throughout, we're working with *qudits* of dimension $d \geq 2$ unless otherwise specified. Since the unitary group $\mathrm{U}(d)$ acts on the Hilbert space of dimension d, it has a corresponding *Haar measure* which is invariant under the action of the group. *Haar random states* sampled proportional to this measure are ubiquitous in quantum information, and essential to define our measurement in Section 3.

To perform the necessary calculations on Haar random states, we need to discuss their moments, and some ancillary concepts.

Definition 7. For integer $k \geq 1$, *k-th moment* of an ensemble $\mathcal{E}$ of quantum states is

$$\mathbb{E}_{|\psi\rangle\sim\mathcal{E}}[|\psi\rangle\langle\psi|^{\otimes k}].$$

An ensemble $\mathcal{E}$ is a *(state) t-design* if the moments $1 \leq k \leq t$ are identical to those of the Haar distribution (see Lemma 10).

Definition 8 (permutation operator)**.** Given a permutation $\pi \in \mathrm{S}_s$ (for $s \geq 1$), define a permutation operator $W_\pi \in \mathbb{C}^{d^s \times d^s}$ such that

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

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

Definition 9 (symmetric subspace)**.** The symmetric subspace of an s-qudit system $(\mathbb{C}^d)^{\otimes s}$ is the subspace invariant under W_π for all $\pi \in \mathrm{S}_s$. We use κ_s to denote its dimension and define $\Pi^{(s)}_{\mathrm{sym}}$ to be the projector onto it (notationally omitting the dependence on d, the dimension of the qudit).

We have two characterizations of the symmetric subspace.

Fact 1. *For all s, $\Pi^{(s)}_{\mathrm{sym}} = \frac{1}{s!}\sum_{\pi\in\mathrm{S}_s} W_\pi$, and $\kappa_s = \binom{s+d-1}{d-1}$.*

The integral of $|\psi\rangle\langle\psi|$ over the Haar measure is known from, e.g., [8].

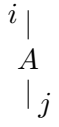
Lemma 10.

$$\kappa_s \int_\psi (|\psi\rangle\langle\psi|)^{\otimes s} \mathrm{d}\psi = \Pi^{(s)}_{\mathrm{sym}} = \frac{1}{s!}\sum_{\pi\in\mathrm{S}_s} W_\pi$$

where $\Pi^{(s)}_{\mathrm{sym}}$ is the projector onto the symmetric subspace and W_π is the operator that permutes s qudits by an s-element permutation π.

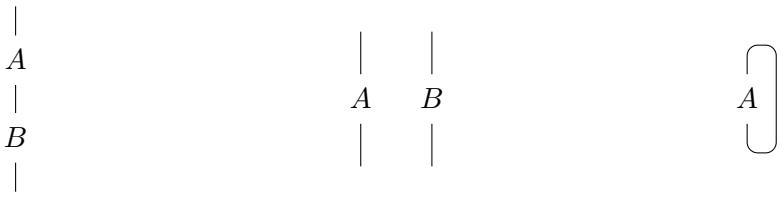
We will often need to compute the (partial) trace of $(A_1 \otimes A_2 \otimes \cdots \otimes A_s)W_\pi$ for some linear operators $A_1, \ldots, A_s \in \mathbb{C}^{d\times d}$. It turns out that there is an extremely useful tensor network based pictorial representation that simplifies these calculations. Let us give a brief introduction to those techniques, though readers may also find more thorough treatments useful [9, 10].

To start, we draw a single d dimensional linear operator $A = \sum_{i,j\in[d]} a_{i,j}|i\rangle\langle j|$ as a tensor block with a leg for the input and output indices for A:

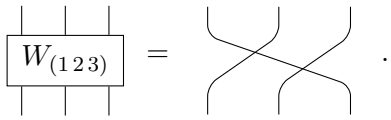


Suppose we have another tensor $B = \sum_{i,j\in[d]} b_{i,j}|i\rangle\langle j|$. We express composition, tensor product, and trace as the following tensor networks:

Composition (AB)	Tensor Product ($A \otimes B$)	Trace ($\mathrm{Tr}(A)$)



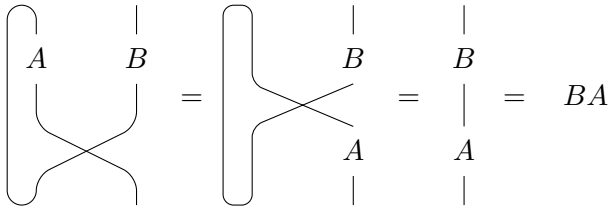
The reason the tensor network picture is particularly nice for dealing with traces of W_π terms is because each W_π term is simply a permutation of wires in the tensor network picture. For example, for a simple cyclic permutation, we have



.

The key feature of tensor networks is that only the topology of the network matters, so we can simplify tensor networks just by moving the elements around. For example, consider a common partial trace that will arise in this paper: $\mathrm{Tr}_1((A \otimes B)W_{(1\,2)})$. Drawing the tensor network, we get

$\mathrm{Tr}_1((A \otimes B)W_{(1\,2)}) \quad =$



BA

where we can push the B tensor through the SWAP and around the trace loop to see that it is composed with A. In other words, we have just shown the identity $\mathrm{Tr}_1((A \otimes B)W_{(1\,2)}) = BA$.

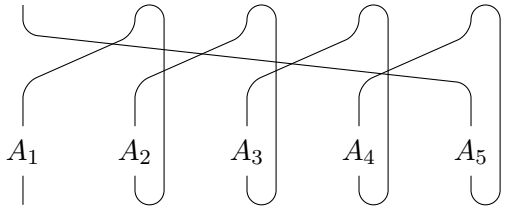
As a generalization, we have the following useful fact:

Fact 2. *Let $n \geq 1$ and $\pi = (1\,2 \cdots n) \in \mathrm{S}_n$. For any $A_1, \ldots, A_n$, we have*

$$\mathrm{Tr}_{-1}(W_\pi(A_1 \otimes A_2 \otimes \cdots \otimes A_n)) = A_n A_{n-1} \cdots A_1,$$

where Tr_{-1} indicates the partial trace of all but the first qudit. Thus, $\mathrm{Tr}(W_\pi(A_1 \otimes A_2 \otimes \cdots \otimes A_n)) = \mathrm{Tr}(A_n A_{n-1} \cdots A_1)$.

Proof. The fact is best seen with a small example. When $n = 5$, for instance, the tensor network diagram is



from which the identity follows. □

3 Joint Measurement Upper Bound

The goal of this section is to prove the following upper bound for the sample complexity of classical shadows for pure states and joint measurements.

Theorem 11.

$$\mathsf{Shadows}(B, \epsilon, \delta) = \mathcal{O}\left(\left(\frac{\sqrt{B}}{\epsilon} + \frac{1}{\epsilon^2}\right) \log \frac{1}{\delta}\right).$$

The proof of Theorem 11 is constructive; given B, ϵ, δ and d, we specify the number of samples and a pair of algorithms $\mathcal{A}_{\text{meas}}$ and $\mathcal{A}_{\text{est}}$ that solve the classical shadows task with that many samples.

In brief, the construction is as follows. We give a measurement $\mathcal{M}_s$ on s copies of ρ, where the outcome of the measurement is a classical description of a pure state Ψ. We apply an affine transformation to the outcome Ψ to produce an unbiased "shadow" estimator: a unital Hermitian matrix $\hat{\rho}$ such that $\mathbb{E}[\hat{\rho}] = \rho$. Increasing the number of samples, s, suppresses the additive error ϵ and the failure probability δ by a factor of $s^{-\mathcal{O}(1)}$. To improve this to an inverse exponential suppression in the failure probability, we repeat the entire procedure $k = \mathcal{O}(\log(\delta^{-1}))$ times and take the median of the batch estimates akin to the *median of means* method [11, 12, 6]. A pseudocode description is given in Algorithms 1 and 2.

Algorithm 1 Algorithm for $\mathcal{A}_{\text{meas}}$ of Theorem 11

Input: Quantum state $\rho^{\otimes N}$, B, ϵ, δ, d.
Output: Classical shadow $\left\{\hat{\rho}^{(i)}\right\}_{i\in[k]}$.

$s \leftarrow \mathcal{O}(\sqrt{B}\epsilon^{-1} + \epsilon^{-2})$ ▷ Samples per batch
$k \leftarrow \lfloor N/s \rfloor$ ▷ Number of batches
for each batch $i = 1, \ldots, k$ **do**
    $\psi_i \leftarrow$ Measure new batch of $\rho^{\otimes s}$ with $\mathcal{M}_s$
    $\hat{\rho}^{(i)} \leftarrow \frac{(d+s)\psi_i - I}{s}$
end for
return $\left\{\hat{\rho}^{(i)}\right\}_{i\in[k]}$

Algorithm 2 Algorithm for $\mathcal{A}_{\text{est}}$ of Theorem 11 and Theorem 29

Input: Classical shadow $\left\{\hat{\rho}^{(i)}\right\}_{i\in[k]}$ and observable O.

$E \leftarrow \text{median}(\text{Tr}(O\hat{\rho}^{(1)}), \ldots, \text{Tr}(O\hat{\rho}^{(k)}))$
return E

We now define our measurement $\mathcal{M}_s$.

Definition 12. The *standard symmetric joint measurement* is a measurement on s qu$d$its. It is defined by the POVM $\mathcal{M}_s = \{A_\psi\}_\psi \cup \{I - \Pi^{(s)}_{\text{sym}}\}$ with elements

$$A_\psi := \kappa_s |\psi\rangle\langle\psi|^{\otimes s} \mathrm{d}\psi,$$

for all d-dimensional pure states, proportional to the Haar measure, plus a "fail" outcome $I - \Pi_{\mathrm{sym}}^{(s)}$ for non-symmetric states.

We will be interested in the setting were ρ is pure[2] and therefore $\rho^{\otimes s}$ is in the symmetric subspace, so we will never see the "fail" outcome—it exists solely to make the POVM sum/integrate to I.

One might be concerned that the standard symmetric joint measurement is constructed from the Haar measure, resulting in a continuum of outcomes. This is technically inconsistent with Definition 2 where the measurement must output a *finite* length bit string. However, it will turn out that our analysis (c.f. Theorem 11) only requires the states that appear in the POVM to form an $(s+2)$-design, where s is the number of samples jointly measured. That is, it suffices to replace the continuous POVM $\mathcal{M}_s$ with a finite POVM $\{\kappa_s p_i |\psi_i\rangle\langle\psi_i|^{\otimes s}\}_i \cup \{I - \Pi_{\mathrm{sym}}^{(s)}\}$ such that $\sum_i p_i |\psi_i\rangle\langle\psi_i|^{\otimes(s+2)} = \int_\psi |\psi\rangle\langle\psi|^{\otimes(s+2)} \mathrm{d}\psi$ where the $p_i \geq 0$ define a finite probability distribution.

For some perspective, consider the independent measurement setting in which $s = 1$. By the above observation, we require the measurement to form a 3-design. Since the set of multi-*qubit* stabilizer states forms a 3-design, we recover the efficient measurement protocol of [6]. That said, our measurements typically involve many copies of the state, resulting in a large s. In such cases, we must use much more complicated constructions of designs (see, e.g., [13, 14, 15]). Nevertheless, these constructions result in a finite POVM that can at least in principle be implemented with a projective measurement using $\mathsf{poly}(d, \log(1/\epsilon))$-many ancillas [16].

3.1 Analysis

After defining the measurement, the estimator, and how many samples we need, the only remaining technical component is to bound the probability of failure. This ultimately comes down to Chebyshev's inequality:

$$\Pr[|\operatorname{Tr}(O\hat{\rho}) - \mathbb{E}[\operatorname{Tr}(O\hat{\rho})]| \geq \epsilon] \leq \frac{\operatorname{Var}(\operatorname{Tr}(O\hat{\rho}))}{\epsilon^2}.$$

Hence, we need to calculate the mean and variance of the random variable $\operatorname{Tr}(O\hat{\rho})$. To be precise, let ρ be a pure state and suppose we measure $\rho^{\otimes s}$ with the standard symmetric joint measurement $\mathcal{M}_s$. Let Ψ be the density matrix random variable for $|\psi\rangle\langle\psi|$, where ψ is the outcome of the measurement. Let's start with the mean:

Lemma 13 (First moment)**.** *For measurement $\mathcal{M}_s$ on pure state $\rho^{\otimes s}$, we have*

$$\mathbb{E}[\Psi] = \frac{I + s\rho}{d + s}.$$

Proof. To start, let's express the expectation as a Haar integral using the definition of $\mathcal{M}_s$:

$$\mathbb{E}[\Psi] = \int \psi \cdot \Pr[\Psi = \psi] = \int \psi \cdot \operatorname{Tr}(A_\psi \rho^{\otimes s}) = \kappa_s \int \psi \cdot \operatorname{Tr}(\psi^{\otimes s} \rho^{\otimes s}) \, \mathrm{d}\psi.$$

[2] This is the first time we use the purity of ρ in our analysis, but certainly not the last.

Using the identity $A\operatorname{Tr}(B) = \operatorname{Tr}_2(A \otimes B)$ for all square matrices A and B, we can apply Lemma 10 to compute the integral above:

$$\mathbb{E}[\Psi] = \kappa_s \int \operatorname{Tr}_{-1}\Big(\psi^{\otimes s+1} \cdot (I \otimes \rho^{\otimes s})\Big)\,\mathrm{d}\psi = \frac{\kappa_s}{\kappa_{s+1}} \frac{1}{(s+1)!} \sum_{\pi \in \mathrm{S}_{s+1}} \operatorname{Tr}_{-1}(W_\pi(I \otimes \rho^{\otimes s})).$$

We attack the right hand side by evaluating $\operatorname{Tr}_{-1}(W_\pi(I \otimes \rho^{\otimes s})$ for each π. In particular, we will show that

$$\operatorname{Tr}_{-1}(W_\pi(I \otimes \rho^{\otimes s})) = \begin{cases} I, & \text{if } \pi(1) = 1, \\ \rho, & \text{otherwise.} \end{cases}$$

To do this, we take the cycle decomposition of π and analyze each cycle separately. Notice that any cycle not involving position 1 is completely traced out and the cycle operator acts on a tensor power of ρ only, so Fact 2 says the trace is $\operatorname{Tr}(\rho^k) = \operatorname{Tr}(\rho) = 1$ (since ρ is pure). Thus, only the cycle through position 1 matters. If $\pi(1) = 1$, then this cycle is trivial, and the result is I. Otherwise, the cycle visits $k \geq 1$ copies of ρ, leading to the product $\rho^k = \rho$.

There are $s!$ permutations which fix 1 (i.e., $\pi(1) = 1$) and hence $s \cdot s!$ which do not, so we conclude that

$$\mathbb{E}[\Psi] = \frac{\kappa_s}{\kappa_{s+1}} \frac{1}{(s+1)!} \sum_{\pi \in \mathrm{S}_{s+1}} \operatorname{Tr}_{-1}(W_\pi(I \otimes \rho^{\otimes s})) = \frac{s! \cdot I + s \cdot s! \cdot \rho}{(d+s)s!} = \frac{I + s\rho}{d+s}.$$

Notice that $\operatorname{Tr}(\mathbb{E}[\Psi]) = \frac{\operatorname{Tr}(I) + s\operatorname{Tr}(\rho)}{d+s} = 1 = \mathbb{E}[\operatorname{Tr}(\Psi)]$, as a sanity check. □

We now turn to the variance calculation, which depends on the second moment of the estimator:

Lemma 14 (Second moment)**.** *For measurement $\mathcal{M}_s$ on pure state $\rho^{\otimes s}$, we have*

$$\mathbb{E}[\Psi \otimes \Psi] = \frac{2}{(d+s)(d+s+1)} \left((I + s\rho)^{\otimes 2} - \frac{s(s+1)}{2} (\rho \otimes \rho) \right) \Pi^{(2)}_{\mathrm{sym}}$$

Proof. As in Lemma 13, we evaluate $\mathbb{E}[\Psi \otimes \Psi]$ as

$$\begin{aligned} \mathbb{E}[\Psi \otimes \Psi] &= \int (\psi \otimes \psi) \cdot \Pr[\Psi = \psi] \\ &= \int (\psi \otimes \psi) \cdot \kappa_s \operatorname{Tr}(\psi^{\otimes s} \rho^{\otimes s})\,\mathrm{d}\psi \\ &= \kappa_s \int \operatorname{Tr}_{-1,2}\Big(\psi^{\otimes s+2} \cdot (I^{\otimes 2} \otimes \rho^{\otimes s})\Big)\,\mathrm{d}\psi \\ &= \frac{\kappa_s}{\kappa_{s+2}} \frac{1}{(s+2)!} \sum_{\pi \in \mathrm{S}_{s+2}} \operatorname{Tr}_{-1,2}(W_\pi(I^{\otimes 2} \otimes \rho^{\otimes s})), \end{aligned}$$

where the partial trace $\operatorname{Tr}_{-1,2}$ now preserves the first two qudits. In Figure 1 and for the special case of $s = 1$, we show a complete derivation of how this trace simplifies using the tensor network notation, which may be useful to some readers before proceeding to the more general proof.

Let us evaluate the sum term-by-term. We divide the permutations into two types: those where 1 and 2 appear in separate cycles (type A) and those where 1 and 2 appear in the same cycle (type B). Consider the type A permutations first:

$$\mathrm{Tr}_{-1,2}(W_\pi(I^{\otimes 2} \otimes \rho^{\otimes s})) = \begin{cases} I \otimes I & \text{if } \pi(1) = 1 \text{ and } \pi(2) = 2 \\ I \otimes \rho & \text{if } \pi(1) = 1 \text{ and } \pi(2) \neq 2 \\ \rho \otimes I & \text{if } \pi(1) \neq 1 \text{ and } \pi(2) = 2 \\ \rho \otimes \rho & \text{if } \pi(1) \neq 1 \text{ and } \pi(2) \neq 2 \end{cases}$$

As before, we end up getting I or ρ for each position, depending on whether 1 and 2 were

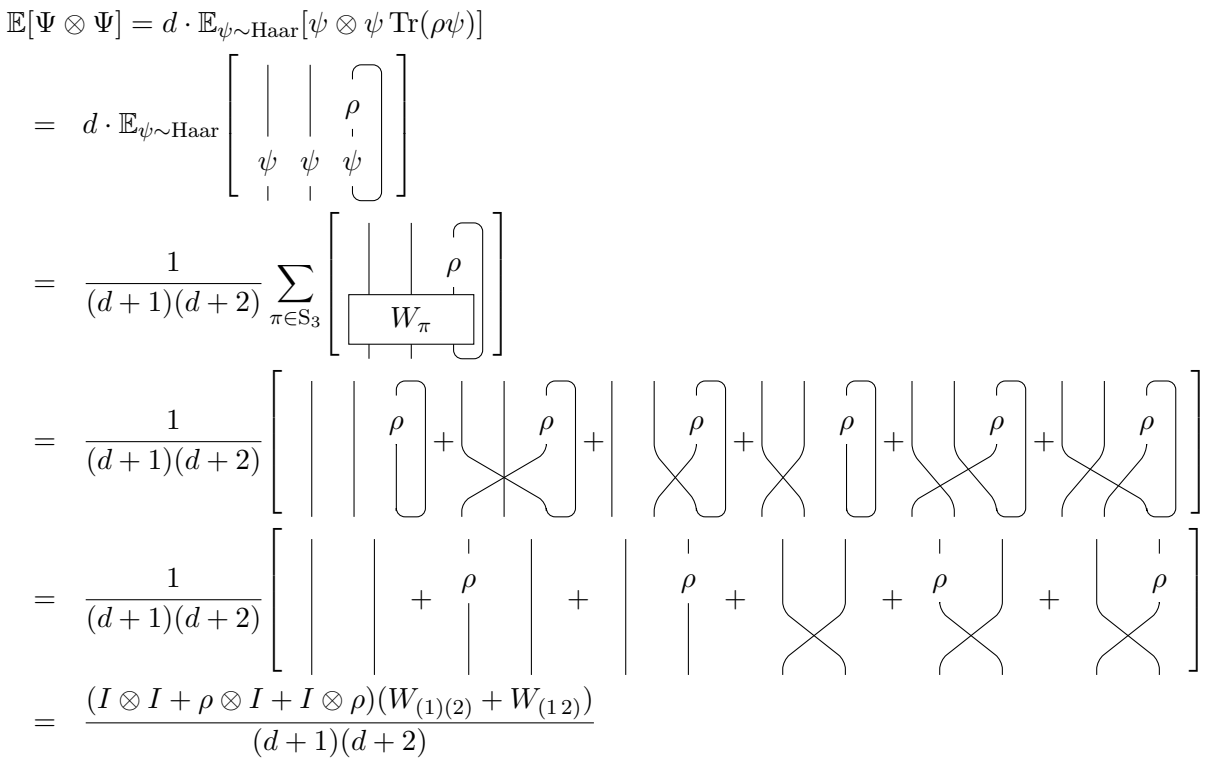


Figure 1: Second moment calculation in the special case $s = 1$.

fixed by the permutation. The combinatorics is similar but not identical: there are $(s+1)!$ permutations which fix 1, and of those, $s!$ fix 2 and $s \cdot s!$ do not. Likewise, $s \cdot s!$ fix 2 but not 1. The remainder of the type A permutations fix neither 1 nor 2, and to count these we need a fact.

Fact 3. *Let $\pi \in \mathrm{S}_n$ be a permutation, and consider $\pi' := (1\,2)\pi$. Then*

$$1 \textit{ and } 2 \textit{ are in distinct cycles of } \pi \iff 1 \textit{ and } 2 \textit{ are in the same cycle of } \pi'.$$

In other words, there is a bijection between A and B permutations, so exactly half of all permutations ($\frac{1}{2}(s+2)!$) are type A, and half are type B. It follows that there are $\frac{s(s-1)}{2} \cdot s!$

type A permutations such that $\pi(1) \neq 1$ and $\pi(2) \neq 2$. Therefore, the overall contribution of the type A permutations is equal to $s!$ times

$$(I \otimes I) + s(I \otimes \rho) + s(\rho \otimes I) + \frac{s(s-1)}{2}(\rho \otimes \rho) = (I + s\rho)^{\otimes 2} - \frac{s(s+1)}{2}\rho^{\otimes 2}.$$

Fortunately, we do not have to repeat this counting argument for the type B permutations. The bijection from Fact 3 decomposes each type B permutation π as $(1\,2)\pi'$ where π' is type A, and so

$$\operatorname{Tr}_{-1,2}(W_\pi(I \otimes \rho^{\otimes s})) = \operatorname{Tr}_{-1,2}(W_{(1\,2)}W_{\pi'}(I \otimes \rho^{\otimes s})) = W_{(1\,2)} \operatorname{Tr}_{-1,2}(W_{\pi'}(I \otimes \rho^{\otimes s})).$$

Therefore, we can multiply our result for type A permutations by $W_{(1)(2)} + W_{(1\,2)} = 2\Pi^{(2)}_{\mathrm{sym}}$ to get the total. The result follows from some careful accounting of the scalar factors. □

Corollary 15. *Let $\hat{\rho} = \frac{(d+s)\Psi - I}{s}$. For any observable $O \in \mathrm{Obs}(B)$ and $\epsilon > 0$, we bound the probability of failure as*

$$\Pr[|\operatorname{Tr}(O\hat{\rho}) - \operatorname{Tr}(O\rho)| \geq \epsilon] \leq \frac{1}{\epsilon^2 s^2}\Big[\operatorname{Tr}(O^2) + 8s \operatorname{Tr}(O^2\rho)\Big].$$

Proof. Our goal will be to compute the mean and variance of the estimate $\operatorname{Tr}(O\hat{\rho})$ in order to apply Chebyshev's inequality. By Lemma 13, we have

$$\mathbb{E}[\hat{\rho}] = \mathbb{E}\left[\frac{(d+s)\Psi - I}{s}\right] = \rho,$$

and so the mean of the estimate is correct: $\mathbb{E}[\operatorname{Tr}(O\hat{\rho})] = \operatorname{Tr}(O\mathbb{E}[\hat{\rho}]) = \operatorname{Tr}(O\rho)$.

To analyze the variance, let us first consider *traceless* observables O, where $\operatorname{Tr}(O) = 0$. As usual, it will be useful to break the variance into a first moment term $\mathbb{E}[\operatorname{Tr}(O\Psi)]$ and a second moment term $\mathbb{E}[\operatorname{Tr}(O\Psi)^2]$:

$$\operatorname{Var}(\operatorname{Tr}(O\hat{\rho})) = \operatorname{Var}\left(\frac{(d+s)\operatorname{Tr}(O\Psi) - \operatorname{Tr}(O)}{s}\right) = \left(\frac{d+s}{s}\right)^2 (\mathbb{E}[\operatorname{Tr}(O\Psi)^2] - \mathbb{E}[\operatorname{Tr}(O\Psi)]^2)$$

Putting aside the s^{-2} factor for now, the (squared) first moment term is

$$(d+s)^2\mathbb{E}[\operatorname{Tr}(O\Psi)]^2 = (d+s)^2\left(\frac{\operatorname{Tr}(O) + s\operatorname{Tr}(O\rho)}{d+s}\right)^2 = s^2 \operatorname{Tr}(O\rho)^2.$$

For the second moment term, we use $\mathbb{E}[\operatorname{Tr}(O\Psi)^2] = \mathbb{E}[\operatorname{Tr}((O \otimes O)(\Psi \otimes \Psi))]$ and Lemma 14 to write

$$\begin{aligned}(d+s)^2\mathbb{E}[\operatorname{Tr}(O\Psi)^2] &= \frac{2(d+s)}{d+s+1}\operatorname{Tr}\left[O^{\otimes 2}\left((I+s\rho)^{\otimes 2} - \frac{s(s+1)}{2}\rho^{\otimes 2}\right)\Pi^{(2)}_{\mathrm{sym}}\right] \\ &\leq \operatorname{Tr}\left[O^{\otimes 2}\left(I^{\otimes 2} + s(I \otimes \rho + \rho \otimes I) - \frac{s(s-1)}{2}\rho^{\otimes 2}\right)(2\Pi^{(2)}_{\mathrm{sym}})\right].\end{aligned}$$

Recall that $(2\Pi^{(2)}_{\mathrm{sym}}) = W_{(1)(2)} + W_{(1\,2)}$, so to simplify, consider the contribution of those two terms:

$$\begin{aligned}W_{(1)(2)}: &\ \operatorname{Tr}(O)^2 + 2s\operatorname{Tr}(O)\operatorname{Tr}(O\rho) + \frac{s(s-1)}{2}\operatorname{Tr}(O\rho)^2 \\ W_{(1\,2)}: &\ \operatorname{Tr}(O^2) + 2s\operatorname{Tr}(O^2\rho) + \frac{s(s-1)}{2}\operatorname{Tr}((O\rho)^2)\end{aligned}$$

In fact, because ρ is pure, we have[3] that $\mathrm{Tr}((O\rho)^2) = \mathrm{Tr}(O\rho)^2$. Combining all of the above, we arrive at a bound for the (scaled) variance of $\mathrm{Tr}(O\Psi)$:

$$(d+s)^2 \operatorname{Var}(\mathrm{Tr}(O\Psi)) \leq \mathrm{Tr}(O^2) + 2s\,\mathrm{Tr}(O^2\rho) - s\,\mathrm{Tr}(O\rho)^2 \leq \mathrm{Tr}(O^2) + 2s\left\|O^2\right\|$$

where the last inequality uses Hölder's inequality.[4]

It follows that $\operatorname{Var}(\mathrm{Tr}(O\hat{\rho}))$ is

$$\operatorname{Var}(\mathrm{Tr}(O\hat{\rho})) \leq \frac{\mathrm{Tr}(O^2) + 2s\|O^2\|}{s^2},$$

for traceless observables. Now suppose O has nonzero trace, and let $O_0 := O - \mathrm{Tr}(O)I/d$ be its traceless part. Naturally, we have

$$\operatorname{Var}(\mathrm{Tr}(O\hat{\rho})) = \operatorname{Var}(\mathrm{Tr}(O_0\hat{\rho})) \leq \frac{\mathrm{Tr}(O_0^2) + 2s\|O_0^2\|}{s^2}.$$

Let's now show that we can bound each of those terms ($\mathrm{Tr}(O_0^2)$ and $\|O_0^2\|$) using functions of the original observable O. First, for the $\mathrm{Tr}(O_0^2)$ term, we have that

$$\mathrm{Tr}(O_0^2) = \mathrm{Tr}(O^2) - \mathrm{Tr}(O)^2/d \leq \mathrm{Tr}(O^2).$$

Next consider the $\|O_0^2\|$ term. We have that $\mathrm{Tr}(O) \leq \|O\|d$, and so the largest eigenvalue (in absolute value) of $O - \mathrm{Tr}(O)I/d$ is at most $2\|O\|$. We get

$$\|O_0^2\| = \|O_0\|^2 \leq (2\|O\|)^2 = 4\|O^2\|.$$

Hence $\operatorname{Var}(\mathrm{Tr}(O\hat{\rho})) \leq \frac{1}{s^2}(\mathrm{Tr}(O)^2 + 8s\|O^2\|)$, and the result follows by Chebyshev's theorem. □

At last, we can prove the main theorem for this section.

Proof of Theorem 11. Consider an arbitrary observable O, and use Corollary 15 to bound the probability a single batch estimate $\hat{\rho}^{(i)}$ is wrong by

$$\Pr[|\,\mathrm{Tr}(O\hat{\rho}^{(i)}) - \mathrm{Tr}(O\rho)| \geq \epsilon] \leq \frac{1}{\epsilon^2 s^2}\left[\mathrm{Tr}(O^2) + 8s\|O^2\|\right] \leq \frac{B + 8s}{\epsilon^2 s^2}.$$

Suppose we want this probability to be less than some constant $p < 1/2$; we leave it to the reader to check that at most $\mathcal{O}(1/(\epsilon^2 p) + \sqrt{B/(\epsilon^2 p)})$ samples suffice, and note that s is chosen accordingly in Algorithm 1.

Recall that our final estimate E is

$$E := \mathrm{median}(\mathrm{Tr}(O\hat{\rho}^{(1)}), \ldots, \mathrm{Tr}(O\hat{\rho}^{(k)})).$$

Assume there are an odd number of batches, so the median is actually some $\mathrm{Tr}(O\hat{\rho}^{(i)})$. If E is a bad estimate, i.e., $|E - \mathrm{Tr}(O\rho)| > \epsilon$ then at least $k/2$ of the batch estimates are wrong: either E and the estimates higher than it, or E and the estimates lower than it.

[3] If $\rho := |\psi\rangle\langle\psi|$, we get $\mathrm{Tr}((O\rho)^2) = \mathrm{Tr}(O|\psi\rangle\langle\psi|O|\psi\rangle\langle\psi|) = \mathrm{Tr}(\langle\psi|O|\psi\rangle\langle\psi|O|\psi\rangle) = \langle\psi|O|\psi\rangle^2 = \mathrm{Tr}((O\rho)^2)$.

[4] Hölder's inequality: For density matrix ρ and observable O, $\mathrm{Tr}(O\rho) \leq \|O\rho\|_1 \leq (\|O\|_\infty \|\rho\|_1) \leq \|O\|_\infty$.

The batches are independent so Chernoff bounds the chance of seeing $\geq k/2$ failures.

$$\Pr[|E - \operatorname{Tr}(O\rho)| \geq \epsilon] \leq \Pr\left[\#\{i : |\operatorname{Tr}(O\hat{\rho}^{(i)}) - \operatorname{Tr}(O\rho)| \geq \epsilon\} \geq \frac{k}{2}\right] \leq \sqrt{4p(1-p)}^{k}$$

Setting this less than the failure probability δ, we have

$$k \geq \frac{\log \delta^{-1}}{\log (4p(1-p))^{-1/2}}.$$

Again, we note that k is set accordingly in Algorithm 1. □

3.2 Discussion

Let us compare this result with the original classical shadows protocol of Huang, Kueng, and Preskill [6]. Their algorithm measures each copy of ρ with $\mathcal{M}_1$[5], producing unbiased single-copy estimates $\hat{\rho}_1, \ldots, \hat{\rho}_s$ for ρ, which are then averaged into a batch estimate $\hat{\rho} = \frac{1}{s}\sum_{i=1}^{s} \hat{\rho}_i$. Given the observable O, the estimate is then $\operatorname{Tr}(O\hat{\rho})$, or the median of several batches, if necessary to reduce the probability of failure.

We have just seen that the variance of a single-copy estimate is $\operatorname{Var}(\operatorname{Tr}(O\hat{\rho}_i)) \leq \operatorname{Tr}(O^2) + \operatorname{Tr}(O^2\rho)$, and averaging s estimates together reduces the variance by a factor of $\frac{1}{s}$. On the other hand, our measurement with $\mathcal{M}_s$ provides an unbiased estimate with variance

$$\operatorname{Var}(\operatorname{Tr}(O\hat{\rho})) \leq \frac{\operatorname{Tr}(O^2)}{s^2} + \frac{\operatorname{Tr}(O^2\rho)}{s}.$$

Since $\operatorname{Tr}(O^2\rho) \leq 1$, we see that the quadratic denominator of $\operatorname{Tr}(O^2)$ (which is the dominant term) is making all the difference.

4 Joint Measurement Lower Bound

Theorem 16. $\mathsf{Shadows}(B, \epsilon, \delta) = \Omega\left(\frac{\sqrt{B}}{\epsilon \log(B+1)} + \frac{\log \delta^{-1}}{\epsilon^2}\right)$ *provided* $B \leq \epsilon d$.

Notice that this bound matches the $\mathcal{O}((\sqrt{B}\epsilon^{-1} + \epsilon^{-2}) \log(\delta^{-1}))$ upper bound up to a $\log(B)$ and a $\log(1/\delta)$ factor. We prove this as two separate lower bounds: $\Omega(\frac{\log \delta^{-1}}{\epsilon^2})$ and $\Omega(\frac{\sqrt{B}}{\epsilon \log(B+1)})$.

The first lower bound ($\Omega(\epsilon^{-2} \log(\delta^{-1}))$ is derived via a reduction from the problem of distinguishing two pure states, ρ_0 and ρ_1, at trace distance 2ϵ from each other. We then use the known performance of the optimal measurement (Helstrom measurement).

The second lower bound is shown via a reduction from a problem in communication complexity known as Boolean Hidden Matching [21]. We will show that any protocol for the classical shadows task implies a protocol for the Boolean Hidden Matching problem, which has known communication complexity lower bounds. These communication lower bounds will

[5]Technically, they use a Clifford unitary instead of a Haar random unitary, presumably for the sake of efficient implementation. However, $\mathcal{M}_1$ will work in place of their measurement, and the analysis is identical since it uses up to third moments of the ensemble of unitaries, which are the same for Clifford vs. Haar random unitaries, i.e., the Cliffords are a 3-design [17, 18, 19, 20].

imply that the classical shadow must contain a significant amount of information. However, Holevo's theorem gives an upper bound on the amount of information gained through measurement. Therefore, in order to successfully complete the classical shadows task, many copies of the unknown state are required.

4.1 $\Omega(\epsilon^{-2}\log(\delta^{-1}))$ lower bound

The proof of the lower bound uses known results relating the trace distance between two states with our ability to distinguish the states by observables or binary measurements. In particular, the maximum gap for the expectation of a positive semi-definite observable is equal to the trace distance between the states:

Lemma 17. *For arbitrary states ρ and σ,*

$$\max_{0\preceq O\preceq I} |\operatorname{Tr}(O\rho) - \operatorname{Tr}(O\sigma)| = \tfrac{1}{2}\|\rho-\sigma\|_1.$$

Furthermore, there is an optimal O satisfying $\operatorname{Tr}(O^2) \le \frac{1}{2}\operatorname{rank}(\rho-\sigma) \le \frac{1}{2}(\operatorname{rank}\rho + \operatorname{rank}\sigma)$.

Proof. Diagonalize $\rho-\sigma$ as $\sum_i \lambda_i |\phi_i\rangle\langle\phi_i|$. Define two positive semi-definite observables, O_+ and O_-.

$$O_+ := \sum_{i:\lambda_i>0} |\phi_i\rangle\langle\phi_i| \qquad\qquad O_- := -\sum_{i:\lambda_i<0} |\phi_i\rangle\langle\phi_i|$$

Clearly $O_+ - O_-$ is a projector onto the eigenvectors of $\rho-\sigma$, so $\operatorname{Tr}((O_+-O_-)(\rho-\sigma)) = \operatorname{Tr}(\rho-\sigma) = 0$, and $\operatorname{rank}(O_+-O_-) = \operatorname{rank}(\rho-\sigma)$. On the other hand,

$$\operatorname{Tr}((O_++O_-)(\rho-\sigma)) = \operatorname{Tr}\left(\sum_{i:\lambda_i\neq 0} (\operatorname{sgn}\lambda_i)\lambda_i|\phi_i\rangle\langle\phi_i|\right) = \sum_{i:\lambda_i\neq 0}|\lambda_i| = \|\rho-\sigma\|_1.$$

It follows that $\operatorname{Tr}(O_+(\rho-\sigma)) = \operatorname{Tr}(O_-(\rho-\sigma)) = \frac{1}{2}\|\rho-\sigma\|_1$. Since O_+ and O_- are orthogonal, the rank of their sum, $\operatorname{rank}(\rho-\sigma)$, is the sum of their ranks. Hence, we can take whichever of O_+ and O_- has rank at most $\frac{1}{2}\operatorname{rank}(\rho-\sigma)$. □

Separately, we know the optimal measurement for distinguishing a uniformly randomly chosen ρ or σ is given by:

Lemma 18 (Helstrom measurement [22])**.** *The optimal measurement for distinguishing states ρ and σ succeeds with probability $\frac{1}{2}+\frac{1}{4}\|\rho-\sigma\|_1$.*

Theorem 19. $\mathsf{Shadows}(1,\epsilon,\delta) = \Omega(\epsilon^{-2}\log(1/\delta))$.

Proof. Let ρ_0 and ρ_1 be pure states with trace distance $\frac{1}{2}\|\rho_0-\rho_1\|_1 = 2\epsilon$. We claim that a protocol for the classical shadows task to ϵ-approximate the expected values of observables of rank 1 with probability of failure at most δ can be used to distinguish states ρ_0,ρ_1 with probability of failure at most δ. First, apply the measurement subroutine to unknown state $\rho_b^{\otimes s}$ to produce the classical shadow. Then, use the observable O from Lemma 17 to estimate $\operatorname{Tr}(O\rho_b)$. If the estimate is closer to $\operatorname{Tr}(O\rho_0)$ then guess $b=0$, otherwise guess $b=1$. Since

the gap $|\operatorname{Tr}(O\rho_0) - \operatorname{Tr}(O\rho_1)| = \frac{1}{2}\|\rho_0 - \rho_1\| = 2\epsilon$, we succeed whenever the gap between the estimate and $\operatorname{Tr}(O\rho_b)$ is less than ϵ.

We can see this classical shadows protocol as a binary measurement distinguishing ρ_0 and ρ_1. Since it succeeds with probability $1-\delta$, the optimal distinguishing measurement from Lemma 18 must do better, so

$$\begin{aligned}(1-2\delta)^2 \leq \frac{1}{4}\|\rho_0^{\otimes s} - \rho_1^{\otimes s}\|_1^2 = 1 - \operatorname{Tr}(\rho_0^{\otimes s}\rho_1^{\otimes s}) = 1 - \operatorname{Tr}(\rho_0\rho_1)^s \\ = 1 - \left(1 - \frac{1}{4}\|\rho_0 - \rho_1\|_1^2\right)^s = 1 - (1-\epsilon^2)^s\end{aligned}$$

where we have used the equation $\frac{1}{2}\|\rho_0 - \rho_1\| = \sqrt{1 - \operatorname{Tr}(\rho_0\rho_1)}$ relating trace distance and fidelity for pure states [23]. Rearranging, we have $(1-\epsilon^2)^s \leq 1 - (1-2\delta)^2 = 4\delta - 4\delta^2 \leq 4\delta$. Taking logs and using $1 - \frac{1}{x} \leq \ln x$ we have

$$-\frac{s\epsilon^2}{1-\epsilon^2} \leq s\log(1-\epsilon^2) \leq \log(4\delta),$$

or equivalently,

$$s \geq \frac{1-\epsilon^2}{\epsilon^2}\log\left(\frac{1}{4\delta}\right) = \Omega(\epsilon^{-2}\log(1/\delta)).$$

□

4.2 $\Omega(\epsilon^{-1}B/\log(B+1))$ lower bound

To prove this lower bound, we leverage the perspective that the classical shadows task is fundamentally a one-way communication problem—recall the setup of the classical shadows task (c.f., Section 1.1) where Melanie measures copies of an unknown state ρ and sends a classical message to Esteban that allows him to estimate the expectation of some observable O on ρ. Intuitively, measuring more copies of ρ means the message will contain more information about ρ. Conversely, if we can prove that Melanie's message must contain a lot of information, we can prove that she must have measured many copies of ρ. In other words, there is a tight correspondence between the sample complexity and one-way classical communication complexity of the classical shadows task.

Formalizing this correspondence is somewhat tricky, so we leave the precise details for later (in particular, Section 4.2.1). However, once this correspondence is established, the high-level structure of the proof is relatively straightforward.

Our starting point is a one-way communication task called "Boolean Hidden Matching". As in the classical shadows task, there are two parties involved in the task: Alice and Bob. Alice has a labeled graph and Bob has a "partial matching" (a collection of vertex-disjoint edges from the graph). Together, these encode a secret bit[6]. Bob doesn't know the labels on the graph, so Alice's goal is to send him a classical message so that he can extract the encoded bit. [21] shows a lower bound on the number of bits that Alice must send to be successful—namely, she must send $\Omega(\sqrt{n/\alpha})$ bits where n is the number of vertices in Alice's graph and α is the fraction of edges in Bob's partial matching.

[6] See Section 4.2.2 for the precise definition.

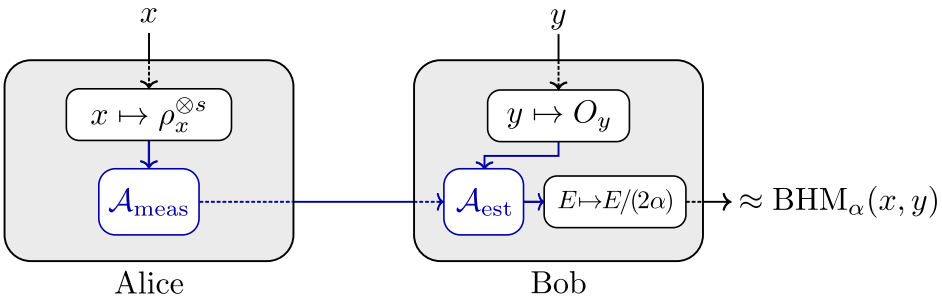


Figure 2: Protocol for Boolean Hidden Matching using classical shadows subroutines (shown in blue). From her input x, Alice prepares $\rho_x^{\otimes s}$, measures, and sends the classical shadow to Bob. From his input y, Bob computes O_y and then estimates $\text{Tr}(O_y \rho_x)$ to accuracy α from the classical shadow that Alice sent to him. He then uses that estimate to answer the Boolean Hidden Matching problem. Correctness follows from the fact that $\text{Tr}(O_y \rho_x) = 2\alpha \cdot \text{BHM}_\alpha(x, y)$. For the details of ρ_x and O_y see Theorem 25.

Our goal will be to take this lower bound for Boolean Hidden Matching and turn it into a lower bound for the classical shadows task. To do this, we create an ensemble of states (corresponding to labeled graphs) and observables (corresponding to partial matchings) such that computing the expected value of an observable with a state solves the Boolean Hidden Matching problem for the corresponding graph and matching. In other words, if Alice and Bob want to solve the Boolean Hidden Matching problem, they can first create the corresponding states and observables, and then use a protocol for classical shadows. We give a depiction of this reduction in Figure 2.

To tie everything together, we appeal to the equivalence between the sample complexity of the classical shadows task and the one-way communication complexity. Namely, Alice must measure a number of copies of her state (roughly) proportional to the number of bits she wants to send Bob. Since we have a lower bound on the number of bits she must send Bob, we have a lower bound on the number of copies she must measure. This completes the proof.

This overall idea draws considerable inspiration from that in the work of Gosset and Smolin for compressing classical descriptions of quantum states [7]. Our proof can be seen as generalization of their techniques.

The remainder of this section is devoted to formalizing the above ideas. Section 4.2.1 introduces one-way communication complexity, culminating in a powerful theorem connecting the number of bits exchanged in a communication protocol and the amount of Shannon information exchanged in the protocol. In Section 4.2.2, we give the formal definition of the Boolean Hidden Matching problem and fill in the missing details from the proof outline above.

4.2.1 One-way communication complexity

Because our proof is based on principles from communication complexity, let's briefly introduce that topic. We are interested in *one-way communication protocols* where two parties—Alice and Bob—are trying to jointly compute some function $f \colon \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. Alice is given some input $x \in \mathcal{X}$ and Bob is given some input $y \in \mathcal{Y}$. Alice's goal is to send a single message $m \in \{0, 1\}^*$ to Bob, so that he can compute $f(x, y)$. Of course, she could choose to send her entire input x, but in many cases it may be possible to communicate fewer bits and still be

successful.

To be precise about the size of the message Alice must send, let X and Y be the random variables (possibly correlated) for the inputs of Alice and Bob, respectively, and let M be the random variable for Alice's message to Bob. Notice that implicit in M is Alice's communication strategy, which may be an arbitrary (randomized) function of her input. Let's start with the easiest setting, where Alice and Bob run deterministic algorithms.

Definition 20 (Deterministic One-Way Communication Complexity)**.** $D_\delta^{(X,Y)}(f)$, the bounded-error deterministic one-way communication complexity of f, is the minimum number of bits that Alice must send to Bob to compute f with at most δ probability of error whenever their inputs are chosen according to the distribution (X, Y).

A natural variant of classical one-way protocols is when Alice and Bob are allowed to run randomized algorithms. There are two settings: private-coin protocols, where Alice and Bob each have access to private random strings; and public-coins protocols, where Alice and Bob also have access to a shared random string along with their private strings. It will not be critical to completely understand the nuances of the various types of protocols for our proof, but we define them in order to precisely state the theorems on which the lower bound rests.

Definition 21 (Randomized One-Way Communication Complexity)**.** $R_\delta(f)$, the bounded-error randomized one-way communication complexity of f, is the minimum number of bits that Alice must send to Bob with a public-coin protocol to compute f over *all* possible inputs with failure probability at most δ.

While randomized strategies may seem more powerful than deterministic strategies, Yao's minimax principle shows that there is always some input distribution for which the randomized and deterministic complexities coincide:

Theorem 22 (Yao's minimax principle [24])**.** $\max_{(X,Y)} D_\delta^{(X,Y)}(f) = R_\delta(f)$.

It turns out that we will eventually be interested in the amount of *information* contained in Alice's message M, not just the length, which is what is measured by the communication complexity. That said, these two quantities are intuitively related—if the information $I(M : X)$ that Alice's message M reveals about her input X is much lower than the number of bits she is communicating, she should be able to send a smaller message and still be successful. The following theorem formalizes this message compression idea:

Theorem 23 ([25])**.** $D_{\Delta+\delta}^{(X,Y)}(f) = 2\Delta^{-1}[\min I(M : X) + O(1)]$ *where the minimization is over all one-way private-coin protocols for f with input distribution (X, Y) and probability of error at most δ.*

In the next section, we will show that classical shadows must also contain a lot of information, which will be the basis of our lower bound.

4.2.2 Classical shadows for Boolean Hidden Matching

Our starting point is a lower bound for the one-way communication complexity for the Boolean Hidden Matching function $\mathrm{BHM}_\alpha \colon \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ which was introduced by Gavinsky, Kempe, Kerenidis, Raz, and de Wolf.

Theorem 24 ([21]). *$R_\delta(\mathrm{BHM}_\alpha) = \Theta(\sqrt{n/\alpha})$ for any constant $\delta < 1/2$.*

Recall that our lower bound technique is to show that a classical shadows strategy with few samples implies a communication protocol for the Boolean Hidden Matching function with low complexity. To do this, let's look more closely at the BHM_α function.

It will be useful describe the function directly as a communication problem with the inputs $x \in \mathcal{X}$ for Alice and $y \in \mathcal{Y}$ for Bob. Alice is given $(0,1)$-assignments for the n vertices of a graph, and Bob is given $(0,1)$-assignments to αn vertex-disjoint edges in the graph. A set of vertex-disjoint edges from a graph is called a *matching*, hence the name of the function. Importantly, the function is only defined on inputs for Alice and Bob that satisfy the following promise: for each edge in the matching, the parity of the connected vertices (from Alice's input) plus Bob's edge bit assignment is some constant $b \in \{0,1\}$. The output of the function is then defined as this bit b.

Formally, for every $n \geq 2$, the Boolean Hidden Matching function $\mathrm{BHM}_\alpha \colon \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ with parameter $\alpha \in (0, 1/4]$ is defined on inputs $\mathcal{X} = \{0,1\}^n$ and $\mathcal{Y} = (\mathbb{N},\mathbb{N})^{\alpha n} \times \{0,1\}^{\alpha n}$ as follows:

Alice:	$x \in \mathcal{X}$.
Bob:	$y = (\mathcal{M}, w) \in \mathcal{Y}$, where $\mathcal{M} = \{(i_1, j_1), \ldots, (i_{\alpha n}, j_{\alpha n})\}$ is a matching and $w \in \{0,1\}^{\alpha n}$ is a bit assignment.
Promise:	There exists $b \in \{0,1\}$ such that $b = x_{i_k} \oplus x_{j_k} \oplus w_k$ for all k.
Output:	$b \in \{0,1\}$

In the setting where α is constant, [21] show that the *quantum* communication complexity of the Boolean Hidden Matching problem is low—Alice only needs to send a $(\log n)$-qubit state. Gosset and Smolin [7] notice that this implies the existence of a set of states and observables whose expectation values give solutions to the Boolean Hidden Matching function. We generalize this observation to the non-constant α setting below:

Theorem 25. *There is a set of states $\{\rho_x \in \mathbb{C}^n\}_{x\in\mathcal{X}}$ and observables $\{O_y \in \mathrm{Obs}(\alpha n)\}_{y\in\mathcal{Y}}$ such that $\mathrm{Tr}(O_y\rho_x) = 2\alpha \cdot \mathrm{BHM}_\alpha(x,y)$. Furthermore, a protocol for the classical shadows task for observables of squared Frobenius norm $B := \alpha n$, estimation accuracy $\epsilon := \alpha$, and failure probability δ implies a one-way private-coin protocol for Boolean Hidden Matching with failure probability δ.*

Proof. Given valid inputs x and $y = (\mathcal{M}, w)$ to the BHM_α function, define the pure state

$$|\psi_x\rangle := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (-1)^{x_i} |i\rangle$$

and the observable

$$O_y := \sum_{k=1}^{\alpha n} \frac{1}{2} (|i_k\rangle - (-1)^{w_k}|j_k\rangle)(\langle i_k| - (-1)^{w_k}\langle j_k|).$$

Notice that $O_y \in \mathrm{Obs}(\alpha n)$ since O_y is a αn-rank projector. Letting $\rho_x := |\psi_x\rangle\langle\psi_x|$, we get

$$\mathrm{Tr}(O_y\rho_x) = \langle\psi_x|O_y|\psi_x\rangle = \frac{1}{n}\sum_{k=1}^{\alpha n}(1 - (-1)^{x_{i_k} \oplus x_{j_k} \oplus w_k}) = 2\alpha b = 2\alpha\,\mathrm{BHM}_\alpha(x,y).$$

In particular, this implies that if E is an α-approximation to $\mathrm{Tr}(O_y\rho_x)$, then

$$|E - \mathrm{Tr}(O_y\rho_x)| < \alpha \implies \left|\frac{E}{2\alpha} - \mathrm{BHM}_\alpha(x,y)\right| < 1/2,$$

or in other words, rounding $E/(2\alpha)$ is equal to $\mathrm{BHM}_\alpha(x,y)$.

We now claim that the existence of these states and observables implies a private-coin one-way protocol for the Boolean Hidden Matching problem (see Figure 2): Suppose we want a protocol for BHM_α with probability of failure at most δ. Let $s = \mathsf{Shadows}(\alpha n, \alpha, \delta)$. On input x, Alice prepares the state $\rho_x^{\otimes s}$, measures it with a valid classical shadows strategy, and sends the resulting classical shadow to Bob. On input y, Bob computes the observable O_y, and then computes an estimate E for $\mathrm{Tr}(O_y\rho_x)$ using the classical shadow sent by Alice. The correctness of the classical shadows strategy implies that E is an α-approximation to $\mathrm{Tr}(O_y\rho_x)$ with probability of failure at most δ. As shown above, Bob can then compute $\mathrm{BHM}_\alpha(x,y)$ with failure probability at most δ by appropriately rounding the estimate. □

Let us now note a key property of the one-way protocol in Theorem 25 for the Boolean Hidden Matching problem. Namely, once Alice prepares $\rho_x^{\otimes s}$, she no longer uses her original input x. Her message (the classical shadow) only depends on her measurement of the state $\rho_x^{\otimes s}$. In particular, if her message is to contain a lot of information about her input x, then Holevo's theorem stipulates that she must be measuring a state of high dimension, or, in other words, s must be large:

Theorem 26 (Holevo [26])**.** *Let Z be the classical outcome of measuring a d-dimensional state drawn from an ensemble $\{\rho_x\}_{x\in\mathcal{X}}$ according to $x \sim X$. Then, $I(X:Z) \le \log d$.*

Naïvely, the states $\rho_x^{\otimes s}$ in Theorem 25 consist of s qudits of dimension n, i.e., it is a space of dimension n^s. However, since each $\rho_x^{\otimes s}$ is invariant under permutation, it belongs to the *symmetric subspace*, which has dimension nearly a factor of $s!$ smaller (see Fact 1).

We are now ready to put all of the pieces of the lower bound together:

Theorem 27. $\mathsf{Shadows}(B,\epsilon) = \Omega\left(\frac{\sqrt{B}}{\epsilon\log(B+1)}\right)$ *provided* $B \le \epsilon d$.

Proof. Using the communication complexity of BHM_α as our starting point, we first show that Alice's message to Bob in every successful protocol for the Boolean Hidden Matching problem must contain a significant amount of information. To show this, note that by Yao's minimax principle (Theorem 22), there exists a distribution (X,Y) such that $R_\delta(\mathrm{BHM}_\alpha) = D_\delta^{(X,Y)}(\mathrm{BHM}_\alpha)$, for any given δ is the probability of error. Theorem 23 lets us upper bound the (deterministic) complexity with mutual information, and Theorem 24 proves a lower bound. Thus,

$$D_\delta^{(X,Y)}(\mathrm{BHM}_\alpha) = O(\min I(M:X) + 1), \text{ and}$$
$$D_\delta^{(X,Y)}(\mathrm{BHM}_\alpha) = \Omega(\sqrt{n/\alpha}),$$

for any constant δ. It follows that $I(M:X) = \Omega(\sqrt{n/\alpha})$ for any one-way private-coin protocol for BHM_α.

Now consider the classical shadows strategy for solving BHM_α as described by Theorem 25, and suppose that Alice measures $s = \mathsf{Shadows}(\alpha n, \alpha)$ copies of ρ_x.[7] Recall that Alice's message M depends only on her measurement of $\rho_X^{\otimes s}$ which has classical outcome Z. We get

$$I(X : M) \leq I(X : Z) \leq \log(\dim \Pi_{\mathrm{sym}}^{(s)}) \leq O(s \log(n/s + 1)),$$

where we have used (in order) the Data Processing Inequality, Holevo's theorem (Theorem 26), the dimension of the symmetric subspace (Fact 1), and the following inequality:

$$\dim \Pi_{\mathrm{sym}}^{(s)} = \binom{n+s-1}{n-1} \leq \binom{n+s}{n} = \binom{n+s}{s} \leq \left(\frac{e(n+s)}{s}\right)^s.$$

Notice that we now have both an upper bound and a lower bound for the mutual information between Alice's input and her message for a one-way protocol for BHM_α with constant error probability. Setting $\epsilon := \alpha$ and $B := \epsilon n$, we have

$$I(X : M) = \Omega(\sqrt{B}/\epsilon), \text{ and}$$
$$I(X : M) = O(s \log(B/(\epsilon s) + 1)).$$

It follows that

$$s = \Omega\left(\frac{\sqrt{B}}{\epsilon \log(B/(\epsilon s) + 1)}\right).$$

Notice that if we substitute any lower bound for s in the RHS of the equation above, then we get a new lower bound for s on the LHS. Unfortunately, plugging in the trivial lower bound ($s \geq 1$) is not very tight. Instead, we will use the $s = \Omega(1/\epsilon^2)$ lower bound from the previous section. To justify this, notice that

$$s = \mathsf{Shadows}(B, \epsilon) \geq \mathsf{Shadows}(1, \epsilon) = \Omega(1/\epsilon^2),$$

where we have used Theorem 19 and the fact that $B \geq 1$ since $\|O\| = 1$. Therefore, we can plug $s = \Omega(1/\epsilon^2)$ into the RHS above to arrive at the following:

$$s = \Omega\left(\frac{\sqrt{B}}{\epsilon \log(B\epsilon + 1)}\right).$$

Assuming $\epsilon \leq 1$, we simplify this to $s = \Omega(\frac{\sqrt{B}}{\epsilon \log(B+1)})$.[8]

Finally, we point out that the construction of Theorem 25 operates in the regime where the states have dimension $d := n$ and the observables are of rank $B = \epsilon d$. One can extend the lower bound to apply to all observables of rank $B \leq \epsilon d$ by embedding the states used in Theorem 25 into a subspace of dimension $n \leq d$ and keeping the observables the same. □

[7] It is possible that Alice could use fewer than $\mathsf{Shadows}(\alpha n, \alpha)$ samples for this specific application to Boolean Hidden Matching, but it will be useful later that s is big enough to estimate a broader class of observables (specifically those used in Theorem 19).

[8] While this simplification apparently makes lower bound weaker for no reason, it doesn't actually effect the overall lower bound. In regimes where $\log(B\epsilon + 1)$ is significantly smaller than $\log(B + 1)$, the additive $1/\epsilon^2$ term in the lower bound becomes dominant.

5 Independent Measurement Upper Bound

Since the global Clifford group acting on qubits is a 3-design, the randomized Clifford measurement classical shadows algorithm of Huang, Kueng, and Preskill [6] can be viewed as simulating independent $\mathcal{M}_1$ measurements on all copies of ρ then, constructing an unbiased estimator from the measurement outcome on each copy. Their result is for independent measurements and general mixed states, but it upper bounds pure states as a special case.

Theorem 28 (Huang, Kueng, Preskill [6])**.** *For all $\epsilon, \delta > 0$,*

$$\mathsf{I\text{-}Shadows}(B, \epsilon, \delta) = \mathcal{O}\left(\frac{B \log(\delta^{-1})}{\epsilon^2}\right).$$

Huang, Kueng, and Preskill also show a matching lower bound, but the hard instances they construct are with states of full rank. We give an independent measurement classical shadows algorithm for pure states which is better in certain parameter regimes (and is no worse).

Theorem 29. *For all $\epsilon, \delta > 0$,*

$$\mathsf{I\text{-}Shadows}(B, \epsilon, \delta) = \mathcal{O}\left(\min\left\{\frac{B}{\epsilon^2}, \frac{\sqrt{Bd}}{\epsilon} + \frac{1}{\epsilon^2}\right\} \log(\delta^{-1})\right).$$

For example, consider the parameter regime in which δ is a constant, $B = d$, and any $\epsilon = o(1)$. Note that this encompasses natural settings such as estimating full-weight Paulis. One can check that (as d grows) the sample complexity given by Theorem 29 is $\mathcal{O}(d/\epsilon + 1/\epsilon^2)$, which is evidently less than $\mathcal{O}(d/\epsilon^2)$, the sample complexity of the Huang-Kueng-Preskill protocol. In general, our approach gives lower sample complexity whenever $\epsilon = o(\sqrt{d/B})$ and $B = \omega(1)$.

As it turns out, our measurement algorithm is the same as the one in [6]—on each copy of ρ, we make an independent measurement with the POVM $\mathcal{M}_1$, which (on multi-qubit systems) can be performed with a random Clifford measurement since we only use third moments of the Haar measure. The difference is in how we construct our estimator for the unknown state. To see this, first let $\Psi_1, \ldots, \Psi_s$ be the Hermitian random variables for the measurement outcomes. Using Lemma 13, notice that $\hat{\rho}_i := (d+1)\Psi_i - I$ is an unbiased estimator for the unknown state, i.e., $\mathbb{E}[\hat{\rho}_i] = \rho$. The average of the $\hat{\rho}_i$'s, i.e.,

$$\hat{X} := \frac{1}{s}\sum_{i=1}^{s} \hat{\rho}_i$$

is effectively the Huang-Kueng-Preskill estimator. Our key observation is that when ρ is pure, $\hat{\rho}_i\hat{\rho}_j$ is also an unbiased estimator of ρ:

$$\mathbb{E}[\hat{\rho}_i\hat{\rho}_j] = \mathbb{E}[\hat{\rho}_i]\mathbb{E}[\hat{\rho}_j] = \rho^2 = \rho.$$

where we have used the independence of the measurements for the first equality and the purity of ρ for the last. In light of this, we consider an estimator $\hat{Y}$ defined to be the average of the $s(s-1)$ quadratic terms where $i \neq j$.

$$\hat{Y} := \frac{1}{s(s-1)}\sum_{i \neq j} \hat{\rho}_i\hat{\rho}_j.$$

To analyze the accuracy of the estimator $\hat{Y}$, we will once again turn to Chebyshev's inequality:

$$\Pr[|\operatorname{Tr}(O\hat{Y}) - \operatorname{Tr}(O\rho)]| \geq \epsilon] \leq \frac{\operatorname{Var}(\operatorname{Tr}(O\hat{Y}))}{\epsilon^2}.$$

Expanding out the variance term using the definition of $\hat{Y}$, we get

$$\operatorname{Var}(\operatorname{Tr}(O\hat{Y})) = \frac{1}{s^2(s-1)^2} \sum_{i \neq j} \sum_{k \neq \ell} \operatorname{Cov}(\operatorname{Tr}(O\hat{\rho}_i\hat{\rho}_j), \operatorname{Tr}(O\hat{\rho}_k\hat{\rho}_\ell)).$$

We need to bound all of these covariance terms to bound the variance. When all indices i, j, k, ℓ are distinct, then the covariance is 0 (by independence). For the other four cases, we rely on corollaries 35, 36, 37, and 38, which we summarize in the following lemma (proof in Appendix B):

Lemma 30. *For each combination of i, j, k, ℓ, $\operatorname{Cov}(\operatorname{Tr}(O\hat{\rho}_i\hat{\rho}_j), \operatorname{Tr}(O\hat{\rho}_k\hat{\rho}_\ell))$ is*

1. *One index matches ($|\{i, j\} \cap \{k, \ell\}| = 1$)*
   - *Match in different positions ($i = \ell$ or $j = k$): $\mathcal{O}(\|O\|^2)$*
   - *Match in same position ($i = k$ or $j = \ell$): $\mathcal{O}(\|O\|^2)$*
2. *Both indices match ($|\{i, j\} \cap \{k, \ell\}| = 2$)*
   - *Order swapped ($i = \ell$ and $j = k$): $\mathcal{O}(Bd)$*
   - *Same order ($i = j$ and $k = \ell$): $\mathcal{O}(Bd)$*

Since $\|O\|^2 \leq 1$, the contribution from the first two terms is extremely small compared to the last term. This gives us the following bound on the variance:

Lemma 31. $\operatorname{Var}(\operatorname{Tr}(O\hat{Y})) = \mathcal{O}(\frac{Bd}{s^2} + \frac{1}{s})$.

Proof. Expand the variance as

$$\operatorname{Var}(\operatorname{Tr}(O\hat{Y})) = \frac{1}{s^2(s-1)^2} \sum_{i \neq j} \sum_{k \neq \ell} \operatorname{Cov}(\operatorname{Tr}(O\hat{\rho}_i\hat{\rho}_j), \operatorname{Tr}(O\hat{\rho}_k\hat{\rho}_\ell)).$$

Using Lemma 30, we account for the contribution of each type of covariance term to get

$$\operatorname{Var}(\operatorname{Tr}(O\hat{Y})) = \mathcal{O}\left(\frac{0 \cdot s^4 + \|O\|^2 \cdot s^3 + \|O\|^2 \cdot s^3 + Bd \cdot s^2 + Bd \cdot s^2}{s^4}\right) = \mathcal{O}\Big(Bd/s^2 + 1/s\Big)$$

where we have used that $\|O\|^2 \leq 1$ and that there are $\mathcal{O}(s^4)$ terms where i, j, k, ℓ are distinct; $\mathcal{O}(s^3)$ terms where exactly one index matches; and $\mathcal{O}(s^2)$ terms where both indices match. □

Putting everything together, we can now prove the claimed sample complexity in Theorem 29.

Proof of Theorem 29. We first point out that the sample complexity of our new estimator is only better (or at least no worse) when $\epsilon \leq \sqrt{B/d}$, so when that does not hold we simply use the original $\hat{X}$ estimator of Huang, Kueng, and Preskill.[9]

Otherwise, we use Algorithm 3 to measure the state and construct several $\hat{Y}$ estimators (line 9, constituting the classical shadow. This shadow is then used in Algorithm 2 for the observable estimation step, which once again uses the median-of-means method where the analysis will be identical to that in the proof of Theorem 11.

It suffices to analyze the variance of the estimator constructed within each batch, which is $\mathcal{O}(Bd/s^2 + 1/s)$ by Lemma 31. To apply Chebyshev's inequality, we need the variance to be at most ϵ^2, which occurs when we have at least $s = \mathcal{O}(\sqrt{Bd}/\epsilon + 1/\epsilon^2)$ samples. □

We simulate the new quadratic estimator as shown in Figure 3. The plots show the empirical variances of the linear and quadratic estimators in a regime where the target observable has large Frobenius norm, namely $B = d$. For the linear estimator, one expects that the variance should decrease linearly with the number of samples. For the quadratic estimator, the variance is $\mathcal{O}(Bd/s^2 + 1/s)$ by Lemma 31. Therefore, whenever Bd/s^2 dominates $1/s$, the variance should decrease *quadratically* in the number of samples. Since the plots are shown on a log-log scale, this should result in a slope of -2. We see this scaling in the graph shown on the right since d is large (the slope of the regression for the linear estimator is $-.998$ and the slope for the quadratic estimator is -1.936). However, for the left graph $Bd = d^2$ is only 64, so we expect that the variance for the quadratic estimator to scale linearly after about $s = 64$ copies. Indeed, one can observe that the lines for the linear and quadratic estimators are essentially parallel after that point. As a final observation, we note that in both graphs the quadratic estimator becomes better than the linear estimator at the point where the variance becomes less than 1. Since the estimate is only useful once the variance is less than 1, one could interpret this as conveying the fact that the quadratic estimator is *always* better than the linear estimator in this particular parameter regime.

6 Open Problems

Our upper and lower bounds almost completely settle the question of sample complexity for the classical shadows task with arbitrary measurements and arbitrary observables, but clearly many questions remain. First, can the remaining discrepancies between our upper and lower bounds be removed? That is, can we get the lower bounds to have the correct dependence on δ (which we conjecture to be $\log(\delta^{-1})$) and remove a $\log B$ factor?

Second, the sample complexity of learning states in the context of tomography is known for all combinations of independent vs. joint measurements, and pure vs. mixed states. Can we characterize the sample complexity of classical shadows as thoroughly? Ref. [6] gives matching bounds for independent measurements and mixed states, and our result gives matching bounds for joint measurements on pure states, but the independent/pure and joint/mixed cases are open. At the very least, we know that in the independent measurement setting that the pure

[9] Actually, it is better to smoothly transition between the estimators $\hat{X}$ and $\hat{Y}$ using a convex combination rather than use a sharp threshold. However, the improvement is only a constant factor and would require computing covariance of linear vs. quadratic terms (e.g., $\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k))$) to justify rigorously. Hence, we simply use one or the other.

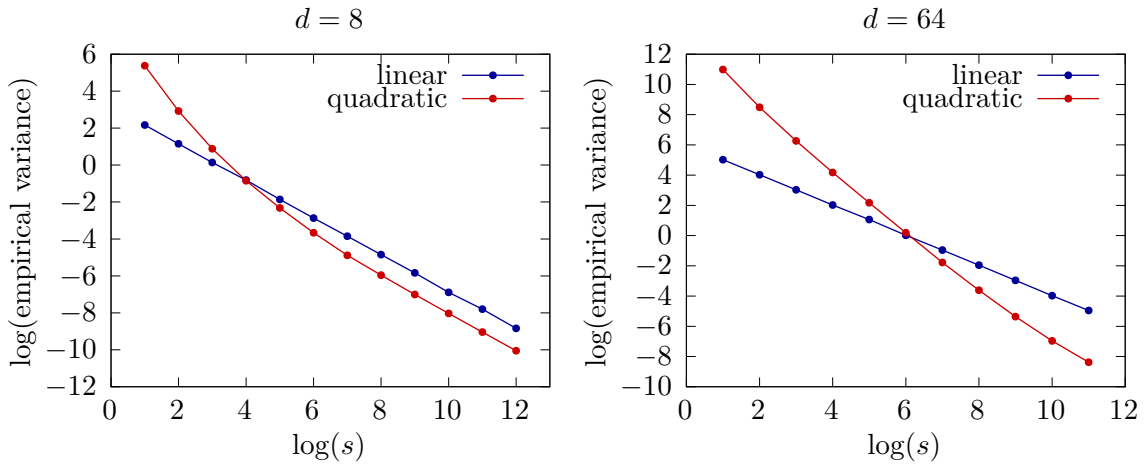


Figure 3: Empirical variances of the linear estimator ($\hat{X}$, the estimator of [6]) and our quadratic estimator ($\hat{Y}$, defined in Section 5) for the classical shadows task using independent measurements. The data confirm our variance calculation in Lemma 31: $\mathcal{O}(\frac{Bd}{s^2} + \frac{1}{s})$. Specifically, the quadratic estimator variance scales inverse-quadratically in s when Bd/s^2 dominates $1/s$ (a slope on the graph of -2), and the linear estimator scales inverse-linearly in s (a slope of -1). Each point in the graphs represent an empirical variance of 10^4 trials of the following procedure: generate a random pure state ρ of dimension $d = 2^n$ and a random full-weight Pauli operator P on n qubits; independently measure s copies of ρ in a random basis to obtain outcomes $\psi_1, \ldots, \psi_s$; compute $\rho_i = (d+1)\psi_i - I$ for all $i \in [s]$; compute estimates $x = \frac{1}{s}\sum_{i=1}^{s} \rho_i$ and $y = \frac{1}{s(s-1)}\sum_{i \neq j} \rho_i \rho_j$; output error from true expectation value: $\mathrm{Tr}(Px) - \mathrm{Tr}(P\rho)$ and $\mathrm{Tr}(Py) - \mathrm{Tr}(P\rho)$. See main text for detailed explanation of the slope of the lines.

state case is different than the mixed state setting since our upper bound in Theorem 29 is smaller than the lower bound in Ref. [6] for some regimes.

As a follow up, tomography bounds are sometimes stated as a function of r, the rank of the unknown state. For example, the sample complexity is $\tilde{\Theta}(rd\epsilon^{-1})$ for joint measurements, capturing the pure state ($r = 1$) and worst-case mixed state ($r = d$) behaviour simultaneously. We believe the sample complexity of the classical shadows task depends smoothly on r, but do not yet have a conjecture.

Third, we do not describe how to concretely implement our large joint measurements. We may replace the continuum of Haar random pure states and elements A_ψ in the POVM $\mathcal{M}_s$ with a concrete $(s+2)$-design, but even then it is not clear how to practically implement the measurement. Alternatively, can we make a similar POVM with a simpler ensemble of states, e.g., t-designs for some $t < s+2$, and update the analysis to achieve an equivalent end result? We note that for our exact measurement, a lower bound on t can be computed by using upper bounds on the number of bits required to describe a state t-design. To see this, notice that the outcome of the POVM $\mathcal{M}_s$ suffices as a compressed description of the state for the purpose of observable estimation. Since we proved a state compression lower bound of $\Omega(\sqrt{B}\epsilon^{-1})$-many bits, the number of bits required to specify the state must be at least this large. In particular, this implies that you could not implement our measurement on n qubits with a 3-design since Clifford states can be specified using $\mathcal{O}(n^2)$-many bits.

Fourth, there is the question of robustness to error or noise. For any classical shadows protocol, we can ask how it behaves when the samples are not exactly of the form $\rho^{\otimes s}$, due

to variation in samples. For our pure state protocols, we are also interested in how fast our algorithms degrade when given mixed states that are close to pure.

Fifth, Ref. [6] introduced two classical shadows protocols known as the random Pauli measurements and the random Clifford measurements schemes. The former schemes targets local observable. The latter scheme, like our protocols, targets observables with low Frobenius norm. These target classes of observables are mutually exclusive and each scheme achieves lower sample complexity with respect to its target class observable. Recent work [27, 28] has also focused on the development of an intermediate scheme that achieves favourable sample complexity scaling in both target classes of observables. All these works focus on general states and independent measurements. Our work identifies an optimal protocol for the low Frobenius norm class of observable in the setting of pure states and joint measurements. So it is natural to consider the pure states and/or joint measurements setting in the context of local observables or the combined class.

Finally, one may consider cubic or higher order generalizations of the quadratic estimator used in the proof of Theorem 29. We leave the analysis of such estimators to future work.

Acknowledgements

We thank Anurag Anshu, Stephen Bartlett, Daniel Burgarth, David Gosset, David Gross and Richard Kueng for useful discussions. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities. HP also acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grants [RGPIN-2019-04198] and [RGPIN-2018-05188].

References

[1] 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 2016). Page 913–925. New York, NY, USA (2016). Association for Computing Machinery.

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

[3] Scott Aaronson. "Shadow tomography of quantum states". In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2018). Page 325–338. New York, NY, USA (2018). Association for Computing Machinery.

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

[5] 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 2022). Pages 574–585. Los Alamitos, CA, USA (2022).

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

[7] David Gosset and John Smolin. "A compressed classical description of quantum states". In 14th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2019). Volume 135 of Leibniz International Proceedings in Informatics (LIPIcs), pages 8:1–8:9. Dagstuhl, Germany (2019). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[8] A J Scott. "Tight informationally complete quantum measurements". Journal of Physics A: Mathematical and General **39**, 13507 (2006).

[9] D. Gross, F. Krahmer, and R. Kueng. "A partial derandomization of PhaseLift using spherical designs". Journal of Fourier Analysis and Applications **21**, 229–266 (2015).

[10] Daniel A. Roberts and Beni Yoshida. "Chaos and complexity by design". Journal of High Energy Physics **2017**, 121 (2017).

[11] G. Lugosi and S. Mendelson. "Mean estimation and regression under heavy-tailed distributions: A survey". Found. Comp. Math. **19**, 1145–1190 (2019).

[12] M. Lerasle. "Lecture notes: Selected topics on robust statistical learning theory" (2019). arXiv:1908.10761.

[13] Bela Bajnok. "Construction of spherical t-designs". Geometriae Dedicata **43**, 167–179 (1992).

[14] A. Hayashi, T. Hashimoto, and M. Horibe. "Reexamination of optimal quantum state estimation of pure states". Phys. Rev. A **72**, 032325 (2005).

[15] Andriy Bondarenko, Danylo Radchenko, and Maryna Viazovska. "Optimal asymptotic bounds for spherical designs". Annals of Mathematics **178**, 443–452 (2013).

[16] Michael A Nielsen and Isaac L Chuang. "Quantum computation and quantum information". Cambridge University Press. (2010).

[17] Zak Webb. "The Clifford group forms a unitary 3-design". Quantum Information and Computation **16**, 1379–1400 (2016).

[18] Huangjun Zhu. "Multiqubit Clifford groups are unitary 3-designs". Physical Review A**96** (2017).

[19] Richard Kueng and David Gross. "Qubit stabilizer states are complex projective 3-designs" (2015). arXiv:1510.02767.

[20] Huangjun Zhu, Richard Kueng, Markus Grassl, and David Gross. "The Clifford group fails gracefully to be a unitary 4-design" (2016). arXiv:1609.08172.

[21] Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald De Wolf. "Exponential separations for one-way quantum communication complexity, with applications to cryptography". In Proceedings of the Annual ACM Symposium on Theory of Computing (STOC 2007). Pages 516–525. New York, NY, USA (2007). Association for Computing Machinery.

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

[23] C.A. Fuchs and J. van de Graaf. "Cryptographic distinguishability measures for quantum-mechanical states". IEEE Transactions on Information Theory **45**, 1216–1227 (1999).

[24] Andrew Chi-Chin Yao. "Probabilistic computations: Toward a unified measure of complexity". In 18th Annual Symposium on Foundations of Computer Science (SFCS 1977). Pages 222–227. IEEE (1977).

[25] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. "The communication complexity of correlation". IEEE Transactions on Information Theory **56**, 438–449 (2010).

[26] Alexander Semenovich Holevo. "Bounds for the quantity of information transmitted by a quantum communication channel". Problemy Peredachi Informatsii **9**, 3–11 (1973). url: http://mi.mathnet.ru/ppi903.

[27] Christian Bertoni, Jonas Haferkamp, Marcel Hinsche, Marios Ioannou, Jens Eisert, and Hakop Pashayan. "Shallow shadows: Expectation estimation using low-depth random Clifford circuits" (2022). arXiv:2209.12924.

[28] Ahmed A. Akhtar, Hong-Ye Hu, and Yi-Zhuang You. "Scalable and flexible classical shadow tomography with tensor networks". Quantum **7**, 1026 (2023).

A Proper Learning Discussion

The classical shadows task is a learning problem: given many samples of a quantum state, we make measurements with the goal of learning the state well enough to approximate arbitrary observables. A learning problem is said to be *proper* if it requires the learned representation to be from the same class—in our case, the class of pure states—as the original object. Our classical shadows algorithm fails to be proper on several counts:

1. The output of the problem is a real number, not the classical description of a quantum state. By definition, the classical shadows task is not a proper learning task.

2. Internally, our algorithm does produce Hermitian matrices $\hat{\rho}^{(i)}$ with trace 1 (the *shadows*), which have the potential to represent a quantum state. However, each such estimate is

   (a) high rank, so it cannot represent a pure state, and

   (b) not positive semi-definite, so it cannot represent a mixed state.

3. The algorithm uses multiple $\hat{\rho}^{(i)}$ shadows, and takes a median of their estimates on each observable. Even if each $\hat{\rho}^{(i)}$ were a quantum state, there may not exist a state exactly consistent with all our medians on a set of observables.

On the other hand, in the limit where the failure probability is extremely small, we can afford to estimate the expectation of the state on an ϵ-cover of the observables. From these values, we can approximate the original state to accuracy ϵ in trace distance. In this regime, the problem is equivalent to tomography, which *is* a proper learning problem.

In this section, we show that any proper learning algorithm for classical shadows would require significantly more samples. Our starting point is a known lower bound for the quantum state tomography question for pure states:

Theorem 32 ([1])**.** *Any quantum algorithm that takes copies of an unknown pure state ρ and outputs a classical estimate $\hat{\rho}$ such that $\|\rho - \hat{\rho}\|_1 \leq \epsilon$ with constant failure probability $\delta < 1$ requires $\Omega(d\epsilon^{-2}/\log(d/\epsilon))$ samples.*

In particular, we show that a proper classical shadows algorithm implies a state tomography algorithm.

Theorem 33. *Suppose there exists a quantum learning algorithm that, given s copies of an unknown d-dimensional state ρ, outputs a classical description of a trace 1, Hermitian PSD matrix $\hat{\rho}$ such that, for all $O \in \mathrm{Obs}(1)$ with failure probability $\delta < 1$:*

$$|\mathrm{Tr}(O\rho) - \mathrm{Tr}(O\hat{\rho})| \leq \epsilon. \tag{2}$$

Then, $s = \tilde{\Omega}(d/\epsilon)$.

Proof. Run the algorithm and feed it $O = \rho$. We have that $\|\rho - \hat{\rho}\|_1 \leq \sqrt{8\epsilon}$ since

$$\begin{aligned}
\epsilon &\geq |\,\mathrm{Tr}(O\rho) - \mathrm{Tr}(O\hat{\rho})| \\
&= |\,\mathrm{Tr}(\rho^2) - \mathrm{Tr}(\rho\hat{\rho})| \\
&= |1 - \mathrm{Tr}(\rho\hat{\rho})| && \rho = \rho^2 \text{ since } \rho \text{ is pure} \\
&= |1 - F(\rho,\hat{\rho})^2| && \mathrm{Tr}(\rho\sigma) = F(\rho,\sigma)^2 \text{ if either is pure} \\
&\geq 1 - F(\rho,\hat{\rho})^2 \\
&\geq \tfrac{1}{8}\|\rho - \hat{\rho}\|_1^2 && \text{Fuchs-van de Graaf inequality}
\end{aligned}$$

where $F(\rho,\sigma) := \mathrm{Tr}(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}})$ is the *fidelity* of ρ and σ.

Hence, if we solve the classical shadows task to error ϵ on this observable, then we have estimated ρ to within Schatten 1-norm distance $\mathcal{O}(\sqrt{\epsilon})$. That is, classical shadows learner is also a quantum tomography algorithm, so we can use the known lower bound from Theorem 32. It follows that a *proper* learning algorithm for the classical shadows task requires $\tilde{\Omega}(d/\epsilon)$ samples. □

In the commonly considered regime where $d \gg \epsilon^{-1}$, the above lower bound is significantly more than both our algorithm (from Section 3) and the original classical shadows algorithm, which use only $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ samples.

B Covariance bounds

The goal of this subsection is to prove Lemma 30, which gives each of the covariance terms $\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k\hat{\rho}_\ell))$ that appears in the expansion of the estimator $\hat{Y}$. Because $\hat{\rho}_j$ appears twice in all of the covariance terms that are non-zero, it will be convenient to explicitly calculate the second moment of $\hat{\rho}_j$.

Lemma 34. *For all j, the second moment of $\hat{\rho}_j$ is*

$$\mathbb{E}[\hat{\rho}_j^{\otimes 2}] = (I \otimes I + I \otimes \rho + \rho \otimes I)\left(W_{(1\,2)} - \frac{2}{d+2}\Pi_{\text{sym}}^{(2)}\right).$$

Proof. Recall that $\hat{\rho}_j$ is obtained through an independent and identical measurement process, so it suffices to analyze a specific $\hat{\rho} := (d+1)\Psi - I$ term. By Lemma 13 and Lemma 14, we compute the first and second moment of Ψ for the special case $s = 1$ as

$$\mathbb{E}[\Psi] = \frac{I+\rho}{d+1} \quad \text{and} \quad \mathbb{E}[\Psi \otimes \Psi] = \frac{I \otimes I + I \otimes \rho + \rho \otimes I}{(d+1)(d+2)}\Big(W_{(1)(2)} + W_{(1\,2)}\Big)$$

Now, expanding out the second moment, we get

$$\begin{aligned}
\mathbb{E}[\hat{\rho}^{\otimes 2}] &= \mathbb{E}[((d+1)\Psi - I)^{\otimes 2}] \\
&= (d+1)^2\mathbb{E}[\Psi \otimes \Psi] - (d+1)(\mathbb{E}[\Psi] \otimes I + I \otimes \mathbb{E}[\Psi]) + I \otimes I \\
&= \frac{d+1}{d+2}(I \otimes I + I \otimes \rho + \rho \otimes I)(W_{(1)(2)} + W_{(1\,2)}) - (I \otimes I + I \otimes \rho + \rho \otimes I) \\
&= (I \otimes I + I \otimes \rho + \rho \otimes I)\left(\frac{(d+1)W_{(1\,2)} - W_{(1)(2)}}{d+2}\right),
\end{aligned}$$

where we recall that $W_{(1)(2)} = I \otimes I$ to obtain the last equality. We arrive at the lemma by writing the final line in terms of the symmetric subspace $\Pi_{\text{sym}}^{(2)} = (W_{(1\,2)} + W_{(1)(2)})/2$. □

Our goal will now be to express all covariance terms in a manner such that we can apply Lemma 34. Since these equations can become quite cumbersome to write out fully, we will often drop the "⊗" symbol in expressions with I and ρ. For example,

$$I \otimes I + \rho \otimes I + I \otimes \rho \to (II + I\rho + \rho I)$$

will be a common abbreviation. We enclose these abbreviations in parentheses when they are multiplied with other terms.

Let's first tackle the covariance terms $\text{Cov}(\text{Tr}(O\hat{\rho}_i\hat{\rho}_j), \text{Tr}(O\hat{\rho}_k\hat{\rho}_\ell))$ where there is only 1 index shared, i.e., $(|\{i,j\} \cap \{k,\ell\}| = 1)$. There are two subcases: a match in different positions ($i = \ell$ or $j = k$); or a match in same position ($i = k$ or $j = \ell$). While the proofs are quite similar, we break them into two separate corollaries.

Corollary 35. *For all distinct i, j, k,*

$$\text{Cov}(\text{Tr}(O\hat{\rho}_i\hat{\rho}_j), \text{Tr}(O\hat{\rho}_j\hat{\rho}_k)) \leq 2\,\text{Tr}(O\rho)^2 \leq 2\|O\|^2.$$

Proof. First, translate covariance to a second moment calculation:

$$\begin{aligned}
\text{Cov}(\text{Tr}(O\hat{\rho}_i\hat{\rho}_j), \text{Tr}(O\hat{\rho}_j\hat{\rho}_k)) &= \mathbb{E}[\text{Tr}(O\hat{\rho}_i\hat{\rho}_j)\,\text{Tr}(O\hat{\rho}_j\hat{\rho}_k)^*] - \text{Tr}(O\rho)\,\text{Tr}(O\rho)^* \\
&= \mathbb{E}[\text{Tr}(O\hat{\rho}_i\hat{\rho}_j)\,\text{Tr}(O\hat{\rho}_k\hat{\rho}_j)] - \text{Tr}(O\rho)^2,
\end{aligned}$$

where the last equality uses that O, ρ_j, ρ_k are Hermitian. The second moment can be further decomposed using independence of $\hat{\rho}_i$, $\hat{\rho}_j$, and $\hat{\rho}_k$.

$$\begin{aligned}
\mathbb{E}[\text{Tr}(O\hat{\rho}_i\hat{\rho}_j)\,\text{Tr}(O\hat{\rho}_k\hat{\rho}_j)] &= \text{Tr}((O \otimes O)\,\mathbb{E}[\hat{\rho}_i\hat{\rho}_j \otimes \hat{\rho}_k\hat{\rho}_j]) \\
&= \text{Tr}((O \otimes O)\,(\mathbb{E}[\hat{\rho}_i] \otimes \mathbb{E}[\hat{\rho}_k])\,\mathbb{E}[\hat{\rho}_j \otimes \hat{\rho}_j])
\end{aligned}$$

Using $\mathbb{E}[\hat{\rho}_i] = \mathbb{E}[\hat{\rho}_k] = \rho$ and Lemma 34, we have

$$\begin{aligned}(\mathbb{E}[\hat{\rho}_i] \otimes \mathbb{E}[\hat{\rho}_k])\mathbb{E}[\hat{\rho}_j \otimes \hat{\rho}_j] &= (\rho\rho)(II + \rho I + I\rho)(W_{(1\,2)} - 2\Pi^{(2)}_{\mathrm{sym}}/(d+2)) \\ &= 3(\rho\rho)(W_{(1\,2)} - 2\Pi^{(2)}_{\mathrm{sym}}/(d+2)),\end{aligned}$$

where we have once again use the purity of ρ. Plugging everything in, we get

$$\begin{aligned}\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_k)) &= 3\,\mathrm{Tr}(O^{\otimes 2}\rho^{\otimes 2}W_{(1\,2)}) - \frac{6}{d+2}\,\mathrm{Tr}(O^{\otimes 2}\rho^{\otimes 2}\Pi^{(2)}_{\mathrm{sym}}) - \mathrm{Tr}(O\rho)^2 \\ &= 3\,\mathrm{Tr}(O\rho)^2 - \frac{6}{d+2}\,\mathrm{Tr}(O\rho)^2 - \mathrm{Tr}(O\rho)^2 \\ &\leq 2\,\mathrm{Tr}(O\rho)^2 \leq 2\|O\|^2.\end{aligned}$$

□

Corollary 36. *For all distinct i, j, k,*

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k\hat{\rho}_j)) \leq 2\,\mathrm{Tr}(O^2\rho) \leq 2\|O\|^2.$$

Proof. The proof is similar to that of Corollary 35. We expand the covariance as

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k\hat{\rho}_j)) = \mathbb{E}[\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j)\,\mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_k)] - \mathrm{Tr}(O\rho)^2$$

and compute the second moment term using independence:

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j)\,\mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_k)] &= \mathrm{Tr}\left((O \otimes O)(\mathbb{E}[\hat{\rho}_i] \otimes I)\mathbb{E}[\hat{\rho}_j \otimes \hat{\rho}_j](I \otimes \mathbb{E}[\hat{\rho}_k])\right) \\ &= \mathrm{Tr}((O \otimes O)(\rho I)(II + \rho I + I\rho)(W_{(1\,2)} - 2\Pi^{(2)}_{\mathrm{sym}}/(d+2))(I\rho))\end{aligned}$$

For the $W_{(1\,2)}$ term, we get

$$\mathrm{Tr}(O^{\otimes 2}(\rho I)(II + I\rho + \rho I)W_{(1\,2)}(I\rho)] = \mathrm{Tr}(O^{\otimes 2}(2\rho I + \rho\rho)W_{(1\,2)}) = 2\,\mathrm{Tr}(O^2\rho) + \mathrm{Tr}(O\rho)^2.$$

The $\Pi^{(2)}_{\mathrm{sym}}$ term subtracts a positive quantity, so we drop it to get an upper bound on covariance.

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k\hat{\rho}_j)) \leq 2\,\mathrm{Tr}(O^2\rho) + \mathrm{Tr}(O\rho)^2 - \mathrm{Tr}(O\rho)^2 = 2\,\mathrm{Tr}(O^2\rho) \leq 2\|O\|^2.$$

□

We now turn to the covariance terms $\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_k\hat{\rho}_\ell))$ which share two indices, i.e., $(|\{i,j\} \cap \{k,\ell\}| = 2)$. Once again, there are two subcases: the order is swapped ($i = \ell$ and $j = k$); or the order is the same ($i = j$ and $k = \ell$).

In both cases, their are terms of the covariance that are proportional to $\mathrm{Tr}(O)$, but interestingly, we cannot assume O is traceless as we have earlier. This is due to the fact that the $\hat{Y}$ estimator does not necessarily have trace 1. Nevertheless, it will turn out that this cannot affect the overall covariance of $\hat{Y}$ too much, as we will show in the following corollaries.

Corollary 37. *For all distinct i, j,*

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_i)) \leq d\,\mathrm{Tr}(O^2) + 6\sqrt{d\,\mathrm{Tr}(O^2)} + \|O\|^2$$

Proof. Using independence, we expand the covariance as

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_i)) = \mathrm{Tr}((O\otimes O)\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i]\mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j]) - \mathrm{Tr}(O\rho)^2.$$

Using Lemma 34, we get an expression for the second moment terms:

$$\begin{aligned}\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i]\mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j] &= (II+\rho I+I\rho)^2(W_{(1\,2)} - 2\Pi^{(2)}_{\mathrm{sym}}/(d+2))^2\\ &= (II+3I\rho+3\rho I+2\rho\rho)(W_{(1)(2)} - 4\Pi^{(2)}_{\mathrm{sym}}/(d+2) + 4\Pi^{(2)}_{\mathrm{sym}}/(d+2)^2)\\ &= (II+3I\rho+3\rho I+2\rho\rho)(W_{(1)(2)} - 4(d+1)\Pi^{(2)}_{\mathrm{sym}}/(d+2)^2).\end{aligned}$$

For the $W_{(1)(2)}$ term in $\mathrm{Tr}((O\otimes O)\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i]\mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j])$, we get

$$\mathrm{Tr}(O^{\otimes 2}(II+3I\rho+3\rho I+2\rho\rho)W_{(1)(2)}) = \mathrm{Tr}(O)^2 + 6\,\mathrm{Tr}(O) + 2\,\mathrm{Tr}(O\rho)^2.$$

For the $\Pi^{(2)}_{\mathrm{sym}}$ term, we get (ignoring the scalar factor)

$$\mathrm{Tr}(O^{\otimes 2}(II+3I\rho+3\rho I+2\rho\rho)\Pi^{(2)}_{\mathrm{sym}}) = \frac{\mathrm{Tr}(O^2) + \mathrm{Tr}(O)^2 + 6\,\mathrm{Tr}(O) + 6\,\mathrm{Tr}(O^2\rho) + 4\,\mathrm{Tr}(O\rho)^2}{2}.$$

Therefore, the covariance $\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_j\hat{\rho}_i))$ is

$$\begin{aligned}&= \mathrm{Tr}(O)^2 + 6\,\mathrm{Tr}(O) + \mathrm{Tr}(O\rho)^2 - \frac{2(d+1)}{(d+2)^2}\Big(\mathrm{Tr}(O^2) + \mathrm{Tr}(O)^2 + 6\,\mathrm{Tr}(O) + 6\,\mathrm{Tr}(O^2\rho) + 4\,\mathrm{Tr}(O\rho)^2\Big)\\ &\le \mathrm{Tr}(O)^2 + 6|\,\mathrm{Tr}(O)| + \mathrm{Tr}(O\rho)^2\\ &\le d\,\mathrm{Tr}(O^2) + 6\sqrt{d\,\mathrm{Tr}(O^2)} + \|O\|^2\end{aligned}$$

where the first inequality drops the negative terms and the last line comes from Cauchy-Schwarz. □

Corollary 38. *For all distinct i, j,*

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j)) = (d+2)\,\mathrm{Tr}(O^2) + (3d-2)\|O\|^2.$$

Proof. We expand the covariance as usual and obtain the following.

$$\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j)) = \mathrm{Tr}((O\otimes O)\mathbb{E}[\hat{\rho}_i\hat{\rho}_j\otimes\hat{\rho}_j\hat{\rho}_i]) - \mathrm{Tr}(O\rho)^2$$

It is not so easy to decompose this into $\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i]$ and $\mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j]$. We introduce a third qudit, and use the following identity:

$$\hat{\rho}_i\hat{\rho}_j\otimes\hat{\rho}_j\hat{\rho}_i = \mathrm{Tr}_3((I\otimes\hat{\rho}_i\otimes\hat{\rho}_i)(\hat{\rho}_j\otimes\hat{\rho}_j\otimes I)W_{(1\,3)(2)}).$$

So, plugging back into the expectation, we get

$$\mathrm{Tr}((O\otimes O)\mathbb{E}[\hat{\rho}_i\hat{\rho}_j\otimes\hat{\rho}_j\hat{\rho}_i]) = \mathrm{Tr}((O\otimes O\otimes I)(I\otimes\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i])(\mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j]\otimes I)W_{(1\,3)(2)}).$$

Once again, we use Lemma 34 to compute

$$\begin{aligned}I\otimes\mathbb{E}[\hat{\rho}_i\otimes\hat{\rho}_i] &= (III+I\rho I+II\rho)\Big(\tfrac{d+1}{d+2}W_{(1)(2\,3)} - \tfrac{1}{d+2}W_{(1)(2)(3)}\Big),\\ \mathbb{E}[\hat{\rho}_j\otimes\hat{\rho}_j]\otimes I &= (III+\rho II+I\rho I)\Big(\tfrac{d+1}{d+2}W_{(1\,2)(3)} - \tfrac{1}{d+2}W_{(1)(2)(3)}\Big).\end{aligned}$$

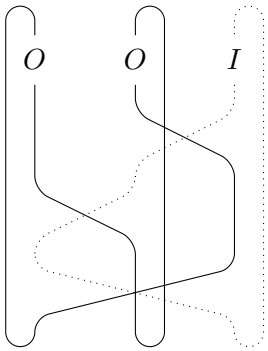


Figure 4: The dominant term in Corollary 38, $\operatorname{Tr}((O \otimes O \otimes I)\, III\, W_{(1)(2\,3)}\, III\, W_{(1\,2)(3)} W_{(1\,3)(2)}) = d \operatorname{Tr}(O^2)$.

Each expectation is a difference of two W_π permutation terms. Therefore, computing the product of the two expectations, we get four W_π terms. It will turn out that the dominant one is the following, where we have taken the $W_{(1)(2\,3)}$ term for $\hat{\rho}_i$ and the $W_{(1\,2)(3)}$ term from $\hat{\rho}_j$ (to visualize the largest contribution from this term, we refer to tensor network picture in Figure 4). Dropping the scalar factor $(d+1)^2/(d+2)^2$ for now, we get

$$\begin{aligned}
&\operatorname{Tr}((O \otimes O \otimes I)(III + I\rho I + II\rho) W_{(1)(2\,3)} (III + \rho II + I\rho I) W_{(1\,2)(3)} W_{(1\,3)(2)}) \\
&= \operatorname{Tr}((O \otimes O \otimes I)(III + I\rho I + II\rho)(III + \rho II + II\rho) W_{(1)(2\,3)} W_{(1\,2)(3)} W_{(1\,3)(2)}) \\
&= \operatorname{Tr}((O \otimes O \otimes I)(III + 3II\rho + I\rho I + \rho II + \rho I\rho + I\rho\rho + \rho\rho I) W_{(1\,2)(3)}) \\
&= \operatorname{Tr}(O^2)(\operatorname{Tr}(I) + 3\operatorname{Tr}(\rho)) + \operatorname{Tr}(O^2\rho)(2\operatorname{Tr}(I) + 2\operatorname{Tr}(\rho)) + \operatorname{Tr}(O\rho)^2 \operatorname{Tr}(I) \\
&= (d+3)\operatorname{Tr}(O^2) + 2(d+1)\operatorname{Tr}(O^2\rho) + d\operatorname{Tr}(O\rho)^2.
\end{aligned}$$

For completeness, let's also compute the other 3 terms (also without their scalar factors):

$$\begin{aligned}
&\operatorname{Tr}((O \otimes O \otimes I)(III + I\rho I + II\rho) W_{(1)(2)(3)} (III + \rho II + I\rho I) W_{(1)(2)(3)} W_{(1\,3)(2)}) \\
&= \operatorname{Tr}((O \otimes O \otimes I)(III + \rho II + 3I\rho I + II\rho + \rho I\rho + \rho\rho I + I\rho\rho) W_{(1\,3)(2)}) \\
&= \operatorname{Tr}(O)^2 + 6\operatorname{Tr}(O)\operatorname{Tr}(O\rho) + 2\operatorname{Tr}(O\rho)^2 \\
&\leq d\operatorname{Tr}(O^2) + 6\sqrt{d}\operatorname{Tr}(O^2) + 2\operatorname{Tr}(O\rho)^2 \leq 6d\operatorname{Tr}(O^2) + 2\operatorname{Tr}(O\rho)^2,
\end{aligned}$$

where inequality comes from Cauchy-Schwarz and the fact that $d \geq 2$. The final two terms turn out to be equal:

$$\begin{aligned}
&\operatorname{Tr}((O \otimes O \otimes I)(III + I\rho I + II\rho) W_{(1)(2\,3)} (III + \rho II + I\rho I) W_{(1)(2)(3)} W_{(1\,3)(2)}) \\
&= \operatorname{Tr}((O \otimes O \otimes I)(III + 3II\rho + I\rho I + \rho II + I\rho\rho + \rho I\rho + \rho\rho I) W_{(1\,2\,3)}) \\
&= \operatorname{Tr}(O^2) + 6\operatorname{Tr}(O^2\rho) + 2\operatorname{Tr}(O\rho)^2
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{Tr}((O \otimes O \otimes I)(III + I\rho I + II\rho) W_{(1)(2)(3)} (III + \rho II + I\rho I) W_{(1)(2\,3)} W_{(1\,3)(2)}) \\
&= \operatorname{Tr}((O \otimes O \otimes I)(III + II\rho + 3I\rho I + \rho II + I\rho\rho + \rho I\rho + \rho\rho I) W_{(1\,2\,3)}) \\
&= \operatorname{Tr}(O^2) + 6\operatorname{Tr}(O^2\rho) + 2\operatorname{Tr}(O\rho)^2
\end{aligned}$$

Notice that these last two terms are non-negative, and so multiplying them by $-(d+1)/(d+2)^2$ makes them non-positive. Since we want to give an upper bound on the covariance, these terms can be dropped. Altogether, and inserting the appropriate constants, we get the following upper bound on the covariance $\mathrm{Cov}(\mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j), \mathrm{Tr}(O\hat{\rho}_i\hat{\rho}_j))$:

$$
\begin{aligned}
&= \mathrm{Tr}((O \otimes O \otimes I)(I \otimes \mathbb{E}[\hat{\rho}_i \otimes \hat{\rho}_i])(\mathbb{E}[\hat{\rho}_j \otimes \hat{\rho}_j] \otimes I)W_{(1\,3)(2)}) \\
&\leq \tfrac{(d+1)^2}{(d+2)^2}((d+3)\,\mathrm{Tr}(O^2) + 2(d+1)\,\mathrm{Tr}(O^2\rho) + d\,\mathrm{Tr}(O\rho)^2) + \frac{(6d\,\mathrm{Tr}(O^2) + 2\,\mathrm{Tr}(O\rho)^2)}{(d+2)^2} - \mathrm{Tr}(O\rho)^2 \\
&\leq \frac{1}{(d+2)^2}((d^3+5d^2+13d+3)\,\mathrm{Tr}(O^2) + (3d^3+7d^2+3d)\|O\|^2) \\
&\leq (d+2)\,\mathrm{Tr}(O^2) + (3d-2)\|O\|^2
\end{aligned}
$$

where we've used once again that $d \geq 2$. □

Algorithm 3 Algorithm for Theorem 29

Input: Quantum state $\rho^{\otimes N}$, B, ϵ, δ, d.
Output: Classical shadow $\left\{\hat{\rho}^{(i)}\right\}_{i\in[k]}$.

$p \leftarrow \mathcal{O}(\log(1/\delta))$ ▷ Number of batches
$s \leftarrow \mathcal{O}(\sqrt{Bd}/\epsilon + 1/\epsilon^2)$ ▷ Samples per batch
$N \leftarrow ps$ ▷ Total number of samples
for each batch $i = 1, \ldots, p$ **do**
    for $j = 1, \ldots, s$ **do**
        $\psi_j^{(i)} \leftarrow$ Measure fresh ρ with $\mathcal{M}_1$
        $\hat{\rho}_j^{(i)} \leftarrow (d+1)\psi_j^{(i)} - I$
    end for
    $\hat{\rho}^{(i)} \leftarrow \frac{1}{s(s-1)} \sum_{j\neq k} \hat{\rho}_j^{(i)} \hat{\rho}_k^{(i)}$
end for
return $\left\{\hat{\rho}^{(i)}\right\}_{i\in[k]}$