On the connection between least squares, regularization, and classical shadows

Zhihui Zhu[1], Joseph M. Lukens[2,3], and Brian T. Kirby[4,5]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA
[2]Research Technology Office and Quantum Collaborative, Arizona State University, Tempe, Arizona 85287, USA
[3]Quantum Information Science Section, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA
[4]DEVCOM Army Research Laboratory, Adelphi, MD 20783, USA
[5]Tulane University, New Orleans, LA 70118, USA

Classical shadows (CS) offer a resource-efficient means to estimate quantum observables, circumventing the need for exhaustive state tomography. Here, we clarify and explore the connection between CS techniques and least squares (LS) and regularized least squares (RLS) methods commonly used in machine learning and data analysis. By formal identification of LS and RLS "shadows" completely analogous to those in CS—namely, point estimators calculated from the empirical frequencies of single measurements—we show that both RLS and CS can be viewed as regularizers for the underdetermined regime, replacing the pseudoinverse with invertible alternatives. Through numerical simulations, we evaluate RLS and CS from three distinct angles: the tradeoff in bias and variance, mismatch between the expected and actual measurement distributions, and the interplay between the number of measurements and number of shots per measurement.

Compared to CS, RLS attains lower variance at the expense of bias, is robust to distribution mismatch, and is more sensitive to the number of shots for a fixed number of state copies—differences that can be understood from the distinct approaches taken to regularization. Conceptually, our integration of LS, RLS, and CS under a unifying "shadow" umbrella aids in advancing the overall picture of CS techniques, while practically our results highlight the tradeoffs intrinsic to these measurement approaches, illuminating the circumstances under which either RLS or CS would be preferred, such as unverified randomness for the former or unbiased estimation for the latter.

Zhihui Zhu: zhu.3440@osu.edu
Joseph M. Lukens: joseph.lukens@asu.edu
Brian T. Kirby: brian.t.kirby4.civ@army.mil

1 Introduction

As experimentally accessible quantum systems continue to increase in size and complexity, methods for characterizing these systems as efficiently as possible have assumed primary importance. One of the leading state characterization approaches is quantum state tomography, which provides a complete density matrix describing a system from which all observable properties can be extracted via classical calculations [1]. However, the exponential scaling in the required number of measurements and the classical computational cost of determining the density matrix most consistent with the given measurement results make state tomography an unrealistic approach for large systems.

The practical challenges associated with full state tomography have spurred the development of various methods for estimating properties and observables of quantum systems without needing to reconstruct the entire density matrix [2]. For example, nonlinear functions of a density matrix, such as various entanglement measures, can be estimated directly without reconstruction but at the cost of requiring multiple copies of a given state and joint measurements between them [3–5]. Further, techniques based on randomized measurements (e.g., application of random uni-

taries to states with fixed measurement bases) have been developed to estimate observables from single-copy systems without requiring state reconstruction. These methods include those that do not incorporate the explicit set of randomized measurements selected into the estimation procedure [6–14] as well as those, such as classical shadows (CS), that do utilize this information and hence still require a shared frame of reference [15–18].

The CS approach to estimating observables of an unreconstructed density matrix is especially attractive due to its experimental simplicity and demonstrated predictive power [15]. The original CS proposal leverages random single-shot measurements to construct "shadows" of a quantum state that then stand-in for a complete density matrix reconstruction for the purpose of calculating observables. Even though the CS density matrix is not even constrained to be positive semidefinite (PSD)—an ostensibly surprising feature critical to its unique scaling behavior [19]—it has been shown to provide accurate estimates for many quantities of interest in a quantum system.

Since its initial development, CS techniques have been studied intensely in various scenarios including—but not limited to—experimental data [18, 20–23], compared to approaches such as Bayesian mean estimation [19], extended to positive operator-valued measures (POVMs) [16, 17] and multiple shots per measurement setting [24], and derandomized to remove the necessity of randomly chosen measurements [25].

From a physical point of view, the original CS proposal is relatively straightforward [15]. The randomness of the measurement procedure induces a depolarizing channel with a simple inversion consisting of subtraction of the identity. Intuitively, since the projections are restricted to a single shot, they naturally act as a noisy channel; they are a low-dimensional projection of a higher-dimensional probability distribution, and the randomness of these projections ensures the noise is isotropic. Importantly from an operational perspective, the inverse of an appropriately chosen random channel can be computed analytically, thereby obviating the need for a computationally intensive inverse calculation.

Here we consider the fundamental causes for the estimation power of CS in light of standard least-squares (LS) and regularized least-squares (RLS) formulations of the same problem. Through a formal derivation of the LS and RLS solutions to a generic quantum measurement scenario, we find that both rely on their own "shadows"—i.e., linear transformations of individual measurement results—which are averaged to obtain the final estimate. As overviewed in Fig. 1 and detailed in Secs. 2 and 3 below, all three techniques (LS, RLS, and CS) follow strikingly similar workflows, differing only in the respective inversion operation in each's shadow formula. Under this viewpoint, these techniques are seen to comprise a general family in which RLS and CS stabilize the LS shadow in the underdetermined regime by replacing the pseudoinverse with invertible and well-conditioned operators. Not only does our work reveal the unifying framework that the idea of "shadows" provides for traditional LS techniques; it also uncovers an profitable interpretation of CS as a regularizer for low-measurement quantum estimation in the tradition of RLS. Collectively, our results contribute to the fundamental understanding of CS while simultaneously offering practical guidance for quantum estimation.

This article is organized as follows. Section 2 introduces the general measurement problem in terms of POVMs and derives the LS solution, expressing the result as an average of LS shadows. Simulations of a five-qubit system reveal high variance and error from the double descent phenomenon, which is mitigated by the RLS and CS stabilization techniques introduced in Sec. 3. Section 4 then compares the advantages and disadvantages of RLS and CS with respect to three specific features: (i) the bias-variance tradeoff, (ii) the impact of misspecified measurement distributions, and (iii) the scaling with reallocations of the number of measurements and the number of shots per measurement. Concluding thoughts appear in Sec. 5.

2 Background: POVM Measurements and LS Estimation

2.1 POVM Measurements

Consider an n-qubit quantum state $\boldsymbol{\rho} \in \mathbb{C}^{D\times D}$ with dimension $D = 2^n$. The probabilistic nature of quantum measurements can be described using POVMs [1]. A POVM is a set of PSD operators

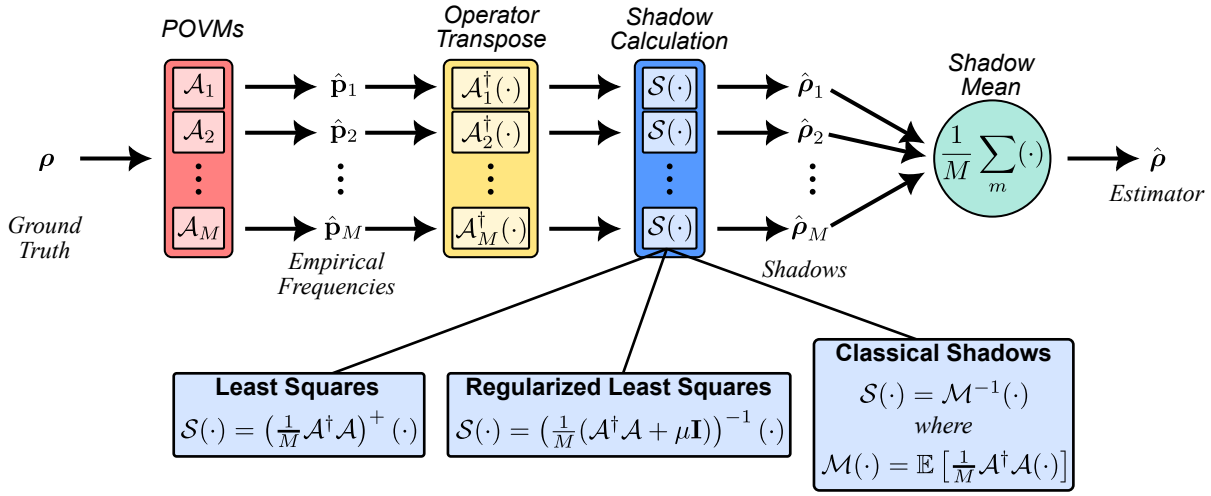


Figure 1: Shadow picture of quantum estimation. POVMs $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_M\} \equiv \mathcal{A}$ are measured via repeated preparation of a ground truth quantum state $\boldsymbol{\rho}$. The observed frequencies for each POVM produce a single shadow state $\widehat{\boldsymbol{\rho}}_m = \mathcal{S}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)$, the collection of which are averaged for the final estimate $\widehat{\boldsymbol{\rho}}$. The only difference between each technique lies in the specific shadow operation chosen: (i) least squares (LS) performs the (pseudo)inverse on the POVMs directly; (ii) regularized least squares (RLS) ensures invertibility through the addition of a term proportional to the identity; and (iii) classical shadows (CS) inverts according to a simulated channel $\mathcal{M}$ defined in expectation over all possible measurement settings.

$\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_K\}$—abbreviated as $\{\boldsymbol{A}_k\}_{k\in[K]}$ with $[K] := \{1, \ldots, K\}$ or simply $\{\boldsymbol{A}_k\}$ when clear from context—such that $\sum_{k=1}^K \boldsymbol{A}_k = \mathbf{I}$. Each POVM element $\boldsymbol{A}_k$ is associated with a possible measurement outcome, and the probability p_k of detecting the k-th outcome when measuring the density operator $\boldsymbol{\rho}$ is given by

$$p_k = \operatorname{tr}(\boldsymbol{A}_k\boldsymbol{\rho}). \tag{1}$$

We can repeat the measurement process L times, observe the k-th outcome f_k number of times, and take the average of the outcomes to generate the empirical frequencies

$$\widehat{p}_k = \frac{f_k}{L}, \ k \in [K]. \tag{2}$$

Collectively, the random variables $f_1, \ldots, f_K$ are characterized by a multinomial distribution with parameters L and $\{p_k\}$. When $L = 1$, the measurements $\{\widehat{p}_k = f_k\}$ form a one-hot vector, or a delta-like distribution, with all entries zero except for one entry being one.

Orthogonal rank-1 POVMs.— A special case common in practice focuses on rank-1 POVMs of the form $\{\boldsymbol{A}_k = \boldsymbol{u}_k\boldsymbol{u}_k^\dagger\}$ with $\boldsymbol{u}_k \in \mathbb{C}^D$ and $\sum_{k=1}^K \boldsymbol{u}_k\boldsymbol{u}_k^\dagger = \mathbf{I}$, where $\dagger$ denotes the Hermitian transpose. We note that in the physics literature when $\boldsymbol{u}$ represents a quantum state it is often represented as a ket $|u\rangle$; however, we adopt vector notation throughout for convenience. When $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{u}_1 & \cdots & \boldsymbol{u}_K\end{bmatrix}^\dagger \in \mathbb{C}^{D\times K}$ further forms an orthonormal basis, in which case $K = D$, the probability p_k can be written as

$$p_k = \operatorname{tr}(\boldsymbol{A}_k\boldsymbol{\rho}) = \boldsymbol{u}_k^\dagger\boldsymbol{\rho}\boldsymbol{u}_k = \boldsymbol{e}_k^\top\left(\boldsymbol{U}\boldsymbol{\rho}\boldsymbol{U}^\dagger\right)\boldsymbol{e}_k, \tag{3}$$

where the last equation implies that the measurement is equivalent to first applying the unitary $\boldsymbol{U}$ to the unknown state $\boldsymbol{\rho} \mapsto \boldsymbol{U}\boldsymbol{\rho}\boldsymbol{U}^\dagger$ (the reason for the Hermitian transpose in the definition $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{u}_1 & \cdots & \boldsymbol{u}_K\end{bmatrix}^\dagger$) and then performing measurements in the canonical (or computational) basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_D$. Both steps can be implemented on a universal quantum computer, though the complexity of synthesizing $\boldsymbol{U}$ by quantum circuits is matrix-dependent. As an aside, we note that rank-1 orthonormal POVMs with $L = 1$ comprised the focus of the original CS proposal [15], although CS extensions to both more generic POVMs [17] and $L > 1$ are possible [24, 26].

POVM ensembles.—As in the case of the above rank-1 POVM, an individual POVM might not achieve informational completeness; therefore measuring with multiple POVMs can be used to acquire a more holistic understanding of the quantum state. For simplicity, consider M POVMs, indexed by $m \in [M]$, where each POVM $\{\boldsymbol{A}_{m,k}\}_{k\in[K]}$ contains the same number of PSD operators K and is probed with L shots, returning the empirical frequencies $\widehat{\boldsymbol{p}}_m$ as described in

Eq. (2), for $m \in [M]$ where the bold notation indicates a vector: $\widehat{\boldsymbol{p}}_m = [f_{m,1} \cdots f_{m,K}]^\top / L = [\widehat{p}_{m,1} \cdots \widehat{p}_{m,K}]^\top$.

To simplify the notation, we collect the probabilities for each POVM $\{\mathrm{tr}\,(\boldsymbol{A}_{m,k}\boldsymbol{\rho})\}$, into a single linear map $\mathcal{A}_m : \mathbb{C}^{D\times D} \to \mathbb{R}^K$ of the form

$$\mathcal{A}_m(\boldsymbol{\rho}) = \begin{bmatrix} \mathrm{tr}\,(\boldsymbol{A}_{m,1}\boldsymbol{\rho}) \\ \vdots \\ \mathrm{tr}\,(\boldsymbol{A}_{m,K}\boldsymbol{\rho}) \end{bmatrix}. \tag{4}$$

If we vectorize $\boldsymbol{\rho}$ and $\boldsymbol{A}_{m,k}$ into $\mathrm{vec}(\boldsymbol{\rho})$ and $\mathrm{vec}(\boldsymbol{A}_k)$ such that $\mathrm{tr}\,(\boldsymbol{A}_k\boldsymbol{\rho}) = (\mathrm{vec}(\boldsymbol{A}_k))^\dagger \mathrm{vec}(\boldsymbol{\rho})$ and define $\boldsymbol{B} = \begin{bmatrix} \mathrm{vec}(\boldsymbol{A}_{m,1}) & \cdots & \mathrm{vec}(\boldsymbol{A}_{m,K}) \end{bmatrix} \in \mathbb{C}^{D^2\times K}$, then $\mathcal{A}_m(\boldsymbol{\rho})$ can be written as matrix-vector product of form

$$\mathcal{A}_m(\boldsymbol{\rho}) = \boldsymbol{B}^\dagger \mathrm{vec}(\boldsymbol{\rho}). \tag{5}$$

Stacking all the empirical frequencies $\{\widehat{\boldsymbol{p}}_m\}$ and the linear operators $\{\mathcal{A}_m\}$ as a single linear map $\mathcal{A} : \mathbb{C}^{D\times D} \to \mathbb{R}^{MK}$, we can write

$$\widehat{\boldsymbol{p}} = \begin{bmatrix} \widehat{\boldsymbol{p}}_1 \\ \vdots \\ \widehat{\boldsymbol{p}}_M \end{bmatrix}, \quad \mathcal{A}(\boldsymbol{\rho}) = \begin{bmatrix} \mathcal{A}_1(\boldsymbol{\rho}) \\ \vdots \\ \mathcal{A}_M(\boldsymbol{\rho}) \end{bmatrix}. \tag{6}$$

It is important to note that no assumptions about informational (tomographic) completeness have been applied in the formalism so far. Specifically, the linear map $\mathcal{A}$ is informationally complete iff $\mathrm{rank}(\mathcal{A}) = D^2$; i.e., we can form exactly D^2 linearly independent operators by linearly combining the set of POVMs [16]. In the regime of interest to CS, $\mathrm{rank}(\mathcal{A}) \ll D^2$ typically holds, although we note that it is possible to formally define a *single informationally complete* POVM so that $\mathrm{rank}(\mathcal{A}) = D^2$ even with $M = 1$—a construction that has been shown valuable for both theoretical analyses [16, 17] and experimental implementation [22] of CS techniques. In this case, any estimator error stems solely from the number of shots L. In our analysis, we do not restrict to informational completeness and in the numerical simulations below consider rank-1 POVMs with $K = D$ outcomes. Nonetheless, the formalism developed applies to any combination of M, K, and L and thus can be explored for any POVMs of potential interest.

2.2 LS Estimation

Without any prior information about $\boldsymbol{\rho}$, we can estimate it from the measurements $\widehat{\boldsymbol{p}}$ by the LS estimator

$$\widehat{\boldsymbol{\rho}} = \underset{\boldsymbol{\rho}'\in\mathbb{C}^{D\times D}}{\arg\min} \left\|\widehat{\boldsymbol{p}} - \mathcal{A}(\boldsymbol{\rho}')\right\|_2^2, \tag{7}$$

by writing $\mathcal{A}(\boldsymbol{\rho}')$ as matrix-vector product as in Eq. (5). While one can explicitly enforce $\boldsymbol{\rho}'$ to be Hermitian and trace one, we will show in Lemma 1 that solutions to Eq. (7) automatically adhere to these properties. Additionally, the solution $\widehat{\boldsymbol{\rho}}$ satisfies the following normal equation

$$\mathcal{A}^\dagger\mathcal{A}(\widehat{\boldsymbol{\rho}}) = \mathcal{A}^\dagger(\widehat{\boldsymbol{p}}), \tag{8}$$

where

$$\begin{aligned} \mathcal{A}^\dagger(\widehat{\boldsymbol{p}}) &= \sum_{m=1}^{M} \mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m) = \sum_{m=1}^{M}\sum_{k=1}^{K} \widehat{p}_{m,k}\boldsymbol{A}_{m,k}, \\ \mathcal{A}^\dagger\mathcal{A}(\widehat{\boldsymbol{\rho}}) &= \sum_{m=1}^{M}\sum_{k=1}^{K} \mathrm{tr}\,(\boldsymbol{A}_{m,k}\widehat{\boldsymbol{\rho}})\,\boldsymbol{A}_{m,k}. \end{aligned} \tag{9}$$

When the ensemble of M POVMs is informationally complete, $\mathcal{A}^\dagger\mathcal{A}$ is invertible and the solution is unique, given by $\widehat{\boldsymbol{\rho}} = \left(\mathcal{A}^\dagger\mathcal{A}\right)^{-1}\left(\mathcal{A}^\dagger(\widehat{\boldsymbol{p}})\right)$. On the contrary, when the M POVMs are not informationally complete, the operator $\mathcal{A}^\dagger\mathcal{A}$ is rank-deficient and the above problem has an infinite number of solutions. Among all possibilities, a common choice is to select the one that has the smallest norm or energy, also known as minimum-norm estimator, which can be obtained by applying the pseudoinverse $(\mathcal{A}^\dagger\mathcal{A})^+$:

$$\begin{aligned} \widehat{\boldsymbol{\rho}} &= \left(\mathcal{A}^\dagger\mathcal{A}\right)^+\left(\mathcal{A}^\dagger(\widehat{\boldsymbol{p}})\right) \\ &= \sum_{m=1}^{M}\left(\mathcal{A}^\dagger\mathcal{A}\right)^+\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right) \\ &= \frac{1}{M}\sum_{m=1}^{M}\underbrace{\left(\frac{1}{M}\mathcal{A}^\dagger\mathcal{A}\right)^+\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)}_{\text{LS shadow}}, \end{aligned} \tag{10}$$

where the significance of defining an LS "shadow" operator is elaborated on below. When the POVMs are informationally complete, $(\mathcal{A}^\dagger\mathcal{A})^+$ becomes $(\mathcal{A}^\dagger\mathcal{A})^{-1}$. Thus, Eq. (10) holds for both informationally complete and incomplete cases.

Lemma 1. *The LS estimator $\widehat{\boldsymbol{\rho}}$ is always Hermitian. Moreover, if $\widehat{\boldsymbol{p}}$ lies in the range space of $\mathcal{A}$, then the LS estimator $\widehat{\boldsymbol{\rho}}$ also has trace* 1.

Proof. Using the two equivalent forms for the pseudoinverse $\mathcal{A}^+ = \left(\mathcal{A}^\dagger\mathcal{A}\right)^+\mathcal{A}^\dagger = \mathcal{A}^\dagger\left(\mathcal{A}\mathcal{A}^\dagger\right)^+$

[27], we can rewrite the LS estimator as $\widehat{\boldsymbol{\rho}} = \mathcal{A}^\dagger\left(\left(\mathcal{A}\mathcal{A}^\dagger\right)^+(\widehat{\boldsymbol{p}})\right)$. Since $\widehat{\boldsymbol{p}}$ is a real vector and $\mathcal{A}\mathcal{A}^\dagger$ is a linear map of $\mathbb{R}^{MK} \to \mathbb{R}^{MK}$, $\left(\mathcal{A}\mathcal{A}^\dagger\right)^+(\widehat{\boldsymbol{p}})$ is also a real vector. As $\widehat{\boldsymbol{\rho}}$ lies in the range space of $\mathcal{A}^\dagger$ that can be written as $\sum_{m=1}^{M}\sum_{k=1}^{K}\alpha_{m,k}\boldsymbol{A}_{m,k}$ with real $\alpha_{m,k}$ and Hermitian $\boldsymbol{A}_{m,k}$, $\widehat{\boldsymbol{\rho}}$ is always Hermitian.

Now if $\widehat{\boldsymbol{p}}$ lies in the range space of $\mathcal{A}$, then the LS estimator $\widehat{\boldsymbol{\rho}}$ satisfies $\mathcal{A}(\widehat{\boldsymbol{\rho}}) = \widehat{\boldsymbol{p}}$, which further implies that $\mathbf{1}^\top\mathcal{A}(\widehat{\boldsymbol{\rho}}) = \mathbf{1}^\top\widehat{\boldsymbol{p}}$. Since $\mathbf{1}^\top\widehat{\boldsymbol{p}} = M$ and $\mathbf{1}^\top\mathcal{A}(\widehat{\boldsymbol{\rho}}) = \sum_{m=1}^{M}\sum_{k=1}^{K}\operatorname{tr}\left(\boldsymbol{A}_{m,k}\widehat{\boldsymbol{\rho}}\right) = M\operatorname{tr}(\widehat{\boldsymbol{\rho}})$, we have $\operatorname{tr}(\widehat{\boldsymbol{\rho}}) = 1$. □

When the operators $\boldsymbol{A}_{1,1},\ldots,\boldsymbol{A}_{M,K}$ are linearly independent, $\widehat{\boldsymbol{p}}$ lies in the range space of $\mathcal{A}$. On the other hand, even when $\widehat{\boldsymbol{p}}$ does not lie in the range space of $\mathcal{A}$—which could happen when $MK > D^2$ and in which case the second half of Lemma 1 does not apply—we have observed in numerical experiments that the LS estimator $\widehat{\boldsymbol{\rho}}$ either has trace 1 or is very close to 1 in practice.

LS shadow.—In anticipation of the CS formalism introduced in Sec. 3.2, we may define $\widehat{\boldsymbol{\rho}}_m := \mathcal{S}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right) = (\frac{1}{M}\mathcal{A}^\dagger\mathcal{A})^+\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)$ the *LS shadow* of $\boldsymbol{\rho}$ associated with measurement m. Unlike CS shadows, these LS shadows can be *biased* estimators of the state $\boldsymbol{\rho}$ as their average $\widehat{\boldsymbol{\rho}}$ biases towards the minimum norm [28]. Yet like CS, the final state estimate $\widehat{\boldsymbol{\rho}}$ is simply the average of the available individual shadows.

Intuitively, one would expect the performance of the LS estimator to improve with more POVM measurements M. We can quantify the agreement of the total estimator $\widehat{\boldsymbol{\rho}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\rho}}_m$ with the ground truth $\boldsymbol{\rho}$ through the Frobenius error $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F$, and the value of any observable $\lambda = \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})$ through the mean squared error (MSE) $\mathbb{E}[(\widehat{\lambda} - \lambda)^2]$, where $\widehat{\lambda} = \operatorname{tr}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\rho}})$.

Computing the error of all estimators with respect to ground truth quantities provides an objective standard by which we can compare all estimation methods in this study. In this vein, it is important to note that the three approaches under consideration (LS, RLS, and CS) correspond to different estimation techniques under a common physical model; i.e., all approaches seek to estimate the same unknown quantum state ρ and assume the same mapping from state to probabilities [Eq. (1)]. In other words, the problem of interest concerns estimation techniques and not model selection, so model identification tools such as the Akaike information criterion [29]—considered in a variety of quantum state estimation contexts [30–33]—do not apply.

On another note, Ref. [15] employed an additional statistical technique, "median of means," to reduce the impact of outliers by partitioning the shadows into several groups and taking the median as the estimate. Incidentally, in recent experimental tests of CS, no significant difference was observed in the performance of the two approaches (mean versus median of means) [18]. Roughly speaking, the median-of-means approach is not designed to reduce the variance of the estimator, but rather obtain a better concentration bound than the sample mean alone [15]. Thus, as this paper focuses on the variance (i.e., MSE) instead of the concentration bound for performance quantification, the sample mean represents the most suitable estimator for our purposes. All that said, we do expect similar phenomena to hold for all comparisons below with the median of means.

For our simulated experiments, we invoke the setup of the original CS proposal [15] with $(K = D)$-outcome, rank-1 POVMs $\{\boldsymbol{A}_{m,1},\ldots,\boldsymbol{A}_{m,D}\}$ defined according to $\boldsymbol{A}_{m,k} = \boldsymbol{u}_{m,k}\boldsymbol{u}_{m,k}^\dagger$, where each $\boldsymbol{U}_m = \left[\boldsymbol{u}_{m,1}\cdots\boldsymbol{u}_{m,D}\right]^\dagger$ is a randomly chosen $D\times D$ unitary matrix. Each POVM measures the state only once (i.e., $L = 1$) so that $\widehat{\boldsymbol{p}}_m$ becomes a one-hot vector, in which case $\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)$ can be rewritten as

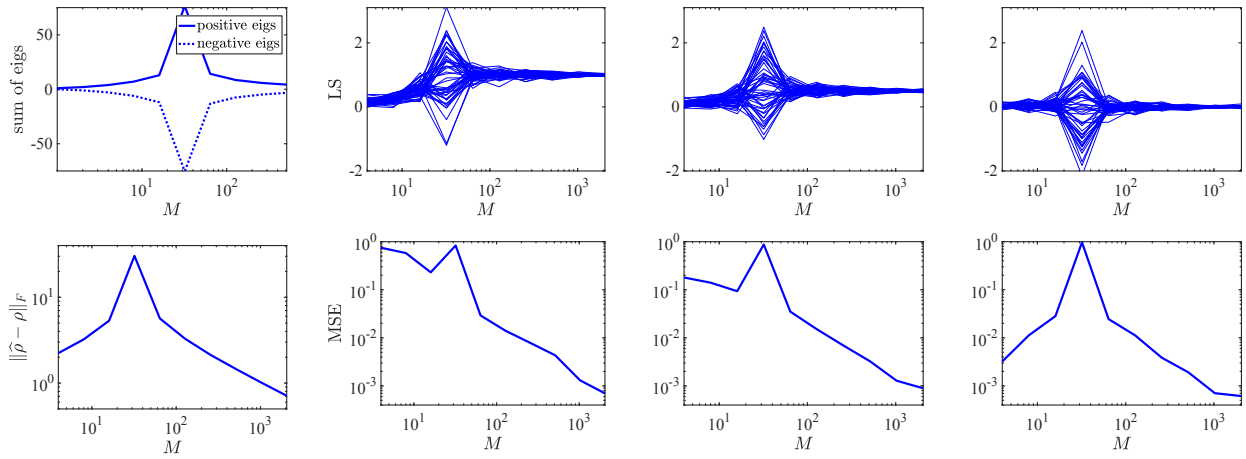
$$\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m) = \sum_{k=1}^{D}\widehat{p}_{m,k}\boldsymbol{u}_{m,k}\boldsymbol{u}_{m,k}^\dagger = (\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)^\dagger. \tag{11}$$

Motivated by our previous study [19], we consider $n = 5$ qubits, Haar-random unitaries, and a fixed ground truth state $\boldsymbol{\rho} = \boldsymbol{e}_0\boldsymbol{e}_0^\dagger$. We focus on three rank-1 observables, $\boldsymbol{\Lambda}_i = \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\dagger, i = 0,1,2$, where

$$\boldsymbol{\phi}_0 = \boldsymbol{e}_0, \boldsymbol{\phi}_1 = \frac{1}{\sqrt{2}}\boldsymbol{e}_0 + \frac{1}{\sqrt{2(D-1)}}\sum_{j=1}^{D-1}\boldsymbol{e}_j, \boldsymbol{\phi}_2 = \boldsymbol{e}_1. \tag{12}$$

These possess ground truth values $\lambda_0 = 1$, $\lambda_1 = 1/2$, and $\lambda_2 = 0$ regardless of dimension D, providing an informative range for exploration.

LS and "double descent" phenomena.—We simulate experiments over 50 independent tri-



(a) $\widehat{\boldsymbol{\rho}}$ (b) $\widehat{\lambda}_0$ (c) $\widehat{\lambda}_1$ (d) $\widehat{\lambda}_2$

Figure 2: Illustration of the performance of the LS shadow for estimating the state $\boldsymbol{\rho}$ and the linear observables $\lambda_i = \operatorname{tr} \boldsymbol{\Lambda}_i \boldsymbol{\rho}$ with $\boldsymbol{\Lambda}_i = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\dagger$, where $\boldsymbol{\phi}_0 = \boldsymbol{e}_0, \boldsymbol{\phi}_1 = \frac{1}{\sqrt{2}} \boldsymbol{e}_0 + \frac{1}{\sqrt{2(D-1)}} \sum_{j=1}^{D-1} \boldsymbol{e}_j, \boldsymbol{\phi}_2 = \boldsymbol{e}_1$: (a) the sum of positive eigenvalues and negative eigenvalues of $\widehat{\boldsymbol{\rho}}$, and $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F$, (b-d) $\widehat{\lambda}_i$ (estimator for λ_i) from 50 independent trials, and the corresponding MSE $(\widehat{\lambda}_i - \lambda_i)^2$ averaged over the 50 trials.

als, in each of which we compute the LS estimator for ten collections of measurements $M \in \{2^2, 2^3, \ldots, 2^{11}\}$. Notably, Fig. 2 shows that the estimation errors for both the state $\boldsymbol{\rho}$ and the linear observables λ_i do not monotonically decrease with the number of POVMs M. In particular, for $\boldsymbol{\rho}$ and λ_2, when M increases, the estimation error first increases in the underdetermined regime ($M < D$), peaks at the interpolation regime when $M \approx D$, and then decreases when entering the overdetermined regime ($M > D$). Here, the three regimes are defined according to the relationship between the total number of outcomes MD and the size of the state D^2, corresponding to the number of equations and parameters in the LS problem [Eq. (7)]. The curves of the estimation error for λ_0 and λ_1 first decrease, then increase, likewise peaking in the interpolating regime ($M \approx D$), and finally decrease with M. This resembles the "double descent" phenomenon observed in deep neural networks: performance first improves, then gets worse, and then improves again with increasing model or data size [34–38]. This phenomenon has been formally studied for linear regression problems under certain statistical models [39–43], and has recently been observed in neural networks for quantum state tomography [44] and polarimetry of vector beams [45].

Roughly speaking, in the interpolating regime $M \approx D$, $(\mathcal{A}^\dagger \mathcal{A})^+$ is unstable (with very large singular values) and the LS estimator $\widehat{\boldsymbol{\rho}}$ becomes highly nonphysical, i.e., it has large negative eigenvalues, resulting in large errors with respect to the ground truth [Fig. 2(a)]. This appears in Fig. 2(b–d) on the estimators $\widehat{\lambda}_i$ as well, which become unstable around the interpolating regime, varying widely between trials. A formal analysis of this phenomenon is beyond the scope of the present investigation and is reserved for future work. Nonetheless, its presence in the context of the LS shadow estimator forms an important springboard for the techniques described in the following section. Both RLS and CS estimation procedures mitigate the issue of double descent by replacing the pseudoinverse $(\mathcal{A}^\dagger \mathcal{A})^+$ in the LS shadow calculation $\mathcal{S}(\cdot)$ with a stabilized alternative: the similarities—and differences—between the RLS and CS solutions in turn reveal an interesting picture of CS estimation as a complementary and computationally efficient "regularizer" for the LS shadow.

3 Stabilizing the LS Estimator

3.1 RLS Estimation

In the underdetermined regime where the number of tested outcomes MD is smaller than the size of the state D^2, regularization has been widely adopted for constraining the resulting solution. In the literature of quantum state tomog-

raphy, regularization or constraint has been exploited for stable, low-measurement reconstruction under the assumption of specific structural features—such as low-rank states [46–51] and matrix product states and operators [52–60]. Without assuming any particular state structure,a common regularization in statistics and machine learning is ℓ_2 regularization, resulting in the so-called RLS or ridge-regression estimator. This has also been widely used in quantum state tomography [61, 62]. The ℓ_2 regularization often leads to a dense solution; in the context of quantum states, it tends to push toward mixed states (lower purity). Specifically, for a given $\mu \geq 0$,

$$\begin{aligned}\widehat{\boldsymbol{\rho}} &= \arg\min_{\boldsymbol{\rho}' \in \mathbb{C}^{D\times D}} \left\{ \left\| \widehat{\boldsymbol{p}} - \mathcal{A}(\boldsymbol{\rho}') \right\|_2^2 + \mu \left\| \boldsymbol{\rho}' \right\|_F^2 \right\} \\ &= (\mathcal{A}^\dagger \mathcal{A} + \mu \mathbf{I})^{-1} \mathcal{A}^\dagger(\widehat{\boldsymbol{p}}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \underbrace{\left(\frac{1}{M} (\mathcal{A}^\dagger \mathcal{A} + \mu \mathbf{I}) \right)^{-1} \mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)}_{\text{RLS shadow}} .\end{aligned} \tag{13}$$

With the introduced ℓ_2 regularization, $\mathcal{A}^\dagger\mathcal{A} + \mu\mathbf{I}$ is invertible and its condition number decreases as μ increases, achieving the purpose of stabilization. Following the discussion of LS shadows, we may call the induced $\widehat{\boldsymbol{\rho}}_m := \mathcal{S}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right) = \left(\frac{1}{M}(\mathcal{A}^\dagger\mathcal{A} + \mu\mathbf{I})\right)^{-1} \mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)$ the "RLS shadow." Similar to LS shadows, these RLS shadows are biased estimators of the ground truth $\boldsymbol{\rho}$.

Choosing a suitable regularization parameter μ requires balancing stabilization of the operator $\mathcal{A}^\dagger\mathcal{A}+\mu\mathbf{I}$ and fitting the measurements, for which techniques like cross-validation or grid search can be used. Recent work [63] shows that an optimal regularization (which may vary with M) that minimizes Frobenius error $\|\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}\|_F$ can mitigate the double descent phenomenon for linear regression under certain statistical assumptions. While such a formal strategy seems impractical here since the state $\boldsymbol{\rho}$ is unknown *a priori*, numerical experiments demonstrate that a small μ can indeed make RLS estimators stable. As shown in Fig. 3, $\mu = 0.1$ leads to monotonically decreasing MSEs for λ_0 and λ_1. While the MSE for λ_2 still peaks in the interpolating regime ($M \approx D$), it is already small (comparable to those for λ_0 and λ_1) and significantly smaller than the one achieved by LS. We could achieve monotonically decreasing estimation error for both λ_2 and $\boldsymbol{\rho}$ through even larger values of μ (e.g., the $\mu = 1$ case in Fig. 3). However, this will substantially bias the estimator $\widehat{\boldsymbol{\rho}}$ to zero, leading to worse estimation for λ_0 and λ_1. In order to balance the performance across different observables, throughout all subsequent experiments, we simply take $\mu = 0.1$ for RLS.

3.2 CS Estimation

CS estimation utilizes a different approach to stabilize the pseudoinverse $(\frac{1}{M}\mathcal{A}^\dagger\mathcal{A})^+$. Assume each POVM $\{\boldsymbol{A}_{m,k}\}_{k\in[K]}$ is independently and randomly generated from an ensemble of POVMs $\mathbb{A}$ according to a certain probability distribution $P(\mathbb{A})$. We may approximate $\frac{1}{M}\mathcal{A}^\dagger\mathcal{A}$ by its expectation

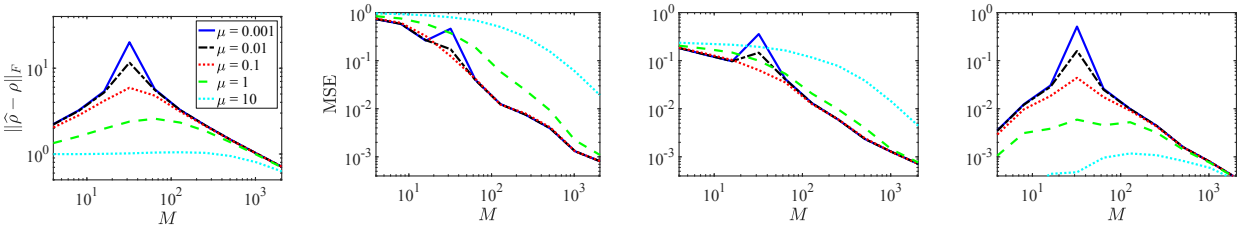
$$\begin{aligned}\mathcal{M}(\boldsymbol{\rho}) &= \mathbb{E}\left[\frac{1}{M}\mathcal{A}^\dagger\mathcal{A}(\boldsymbol{\rho})\right] \\ &= \mathbb{E}_{\{\boldsymbol{A}_k\}\sim P(\mathbb{A})}\left[\sum_{k=1}^{K} \operatorname{tr}\left(\boldsymbol{A}_k\boldsymbol{\rho}\right)\boldsymbol{A}_k\right],\end{aligned} \tag{14}$$

where $\{\boldsymbol{A}_k\}$ represents a random POVM generated from the ensemble of POVMs $\mathbb{A}$ according to the probability distribution $P(\mathbb{A})$. Here $\mathcal{M}$ is called the quantum channel. If $\mathbb{A}$ is tomographically complete, then $\mathbb{E}[\mathcal{A}^\dagger\mathcal{A}]$ is full rank and invertible. The shadow in CS introduced in Ref. [15] can then be defined by replacing $\frac{1}{M}\mathcal{A}^\dagger\mathcal{A}$ in Eq. (10) with its expectation $\mathbb{E}[\mathcal{A}^\dagger\mathcal{A}(\boldsymbol{\rho})]$, i.e.,

$$\widehat{\boldsymbol{\rho}} = \frac{1}{M}\sum_{m=1}^{M} \underbrace{\mathcal{M}^{-1}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)}_{\text{CS shadow}}, \tag{15}$$

so that we obtain the "CS shadow" $\widehat{\boldsymbol{\rho}}_m := \mathcal{S}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right) = \mathcal{M}^{-1}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)$, for which the original CS proposal is named. (We acknowledge the inherent repetitiveness of the term "CS shadow"—"classical shadows shadow"—but adopt it for consistency with LS shadow and RLS shadows.)

As $\mathcal{M}$ defined in Eq. (14) is a linear operator, its inverse $\mathcal{M}^{-1}$ is also linear. These CS shadows are independent and unbiased estimators of $\boldsymbol{\rho}$. Specifically, noting that the randomness of each shadow comes from two sources—the randomly selected POVM $\mathcal{A}_m$ and the random experimental outcome $\widehat{\boldsymbol{p}}_m$—we can take the expectation to obtain



(a) $\widehat{\boldsymbol{\rho}}$ (b) $\widehat{\lambda}_0$ (c) $\widehat{\lambda}_1$ (d) $\widehat{\lambda}_2$

Figure 3: Illustration of the performance of the RLS shadow with different regularization parameter μ for estimating the state $\boldsymbol{\rho}$ and the three linear observables as in Fig. 2.

$$\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\mathcal{M}^{-1}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)\right] = \mathcal{M}^{-1}\underbrace{\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\sum_{k=1}^K \widehat{p}_{m,k}\boldsymbol{A}_{m,k}\right)\right]}_{\mathcal{M}(\boldsymbol{\rho})} = \boldsymbol{\rho}, \tag{16}$$

where $\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\sum_{k=1}^K \widehat{p}_{m,k}\boldsymbol{A}_{m,k}\right] = \mathcal{M}(\boldsymbol{\rho})$ can be obtained by noting the conditional expectation $\mathbb{E}_{\widehat{\boldsymbol{p}}_m}[\widehat{p}_{m,k}\boldsymbol{A}_{m,k} \mid \{\boldsymbol{A}_{m,k}\}] = \mathrm{tr}(\boldsymbol{A}_{m,k}\boldsymbol{\rho})\boldsymbol{A}_k$ according to the Born rule [see Eqs. (1) and (2)].

At this point it is useful to pause and compare the three estimation approaches introduced and analyzed so far: LS, RLS, and CS. As articulated in Fig. 1 and Eqs. (10, 13, 15), each approach produces an estimate that can be viewed as an average over discrete shadows $\widehat{\boldsymbol{\rho}}_m = \mathcal{S}\left(\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)\right)$ each corresponding to one of the M POVMs measured. Whereas LS computes each shadow by direct inversion of the total collection of measurements, both RLS and CS modify this procedure, through regularization and quantum channel inversion, respectively. This observation already offers interesting insights into CS features.

In our opinion, one of the initially most surprising aspects of CS lies in the way it treats the measurement operators post-experiment. Although the POVMs $\mathcal{A}_m$ are selected at random during the measurement process in the canonical CS example, they are known to the user *a posteriori* through the complete collection $\mathcal{A}$. Yet this knowledge is intentionally ignored in the CS shadow operation $\mathcal{S}(\cdot)$; the inversion is instead performed on the *a priori* quantum channel with completely random measurements—an essentially "fictitious" quantum channel from the perspective of the completed experiment. In the light of RLS, however, this channel selection acquires a more intuitive explanation in terms of stabilization: like RLS, CS allows for well-conditioned inversion under any set of measurements, opening the opportunity to improve stability in the estimation procedure and qualitatively accounting for the rigorous information-theoretic bounds it attains [15].

CS with rank-1 POVMs.— The quantum channel $\mathcal{M}$ in Eq. (14) depends on the POVM ensemble and the corresponding sampling distribution. Consider rank-1 POVMs of the form $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d\}$ with $\boldsymbol{A}_k = \boldsymbol{u}_k\boldsymbol{u}_k^\dagger$, where each $\boldsymbol{U} = \begin{bmatrix}\boldsymbol{u}_1 & \cdots & \boldsymbol{u}_d\end{bmatrix}^\dagger$ is randomly chosen from an ensemble of $D \times D$ unitary matrices $\mathbb{U}$. Various unitary ensembles have been explored in prior studies, including the local and global Clifford ensembles [15], fermionic Gaussian unitaries [64], chaotic Hamiltonian evolutions [65], locally scrambled unitary ensembles [65], and Pauli-invariant unitary ensembles [66], for which explicit formulas for the quantum channel $\mathcal{M}$ and its inverse $\mathcal{M}^{-1}$ exist. For instance, if $\mathbb{U}$ is the full unitary group and each unitary matrix $\boldsymbol{U}$ is sampled independently according to the Haar measure on $\mathbb{U}$, the quantum channel $\mathcal{M}$ and its inverse $\mathcal{M}$ can be computed as

$$\begin{aligned}\mathcal{M}(\boldsymbol{\rho}) &= \mathbb{E}_{\boldsymbol{U}}\left[\sum_{k=1}^D \left(\boldsymbol{u}_k^\dagger\boldsymbol{\rho}\boldsymbol{u}_k\right)\boldsymbol{u}_k\boldsymbol{u}_k^\dagger\right] \\ &= \frac{1}{D+1}\boldsymbol{\rho} + \frac{\mathrm{tr}(\boldsymbol{\rho})}{D+1}\mathbf{I},\end{aligned} \tag{17}$$

Table 1: Comparison of RLS and CS estimators.

	distribution independent	computational cost	bias	variance
LS	✓	high	high	high
RLS	✓	high	high	low
CS	✗	low	low (zero)	high

$$\mathcal{M}^{-1}(\boldsymbol{\rho}) = (D+1)\boldsymbol{\rho} - \mathrm{tr}(\boldsymbol{\rho})\mathbf{I}. \tag{18}$$

Plugging this explicit formula and the expression of $\mathcal{A}_m^\dagger(\widehat{\boldsymbol{p}}_m)$ in Eq. (11) into Eq. (13), we can further simplify the CS shadow [15]:

$$\begin{aligned}\widehat{\boldsymbol{\rho}}_m &= (D+1)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)^\dagger \\ &\quad - \mathrm{tr}((\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)^\dagger)\mathbf{I} \\ &= (D+1)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)^\dagger - \mathbf{I},\end{aligned} \tag{19}$$

Since $(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)(\boldsymbol{U}_m^\dagger\widehat{\boldsymbol{p}}_m)^\dagger$ is rank-1 with its only non-zero eigenvalue equal to 1, each classical shadow satisfies $\mathrm{tr}(\widehat{\boldsymbol{\rho}}_m) = 1$, through the summation of one positive eigenvalue equal to D and $D-1$ negative eigenvalues equal to -1.

4 Comparing RLS and CS Stabilization Techniques

4.1 Overview

As introduced and discussed in the previous section, both RLS and CS invoke "fictitious" quantum channels to stabilize the inverse operation in computing each shadow $\widehat{\rho}_m$. Nevertheless, they do so in significantly different ways, leading to distinct advantages and disadvantages in specific use cases. To examine these aspects further, we test and compare RLS and CS methods in the estimation of quantum observables from three perspectives, each of which leverages targeted numerical simulations to reveal the important behaviors of interest. We summarize the three features below and in Table 1, and the relevant numerical simulations follow in the subsequent subsections.

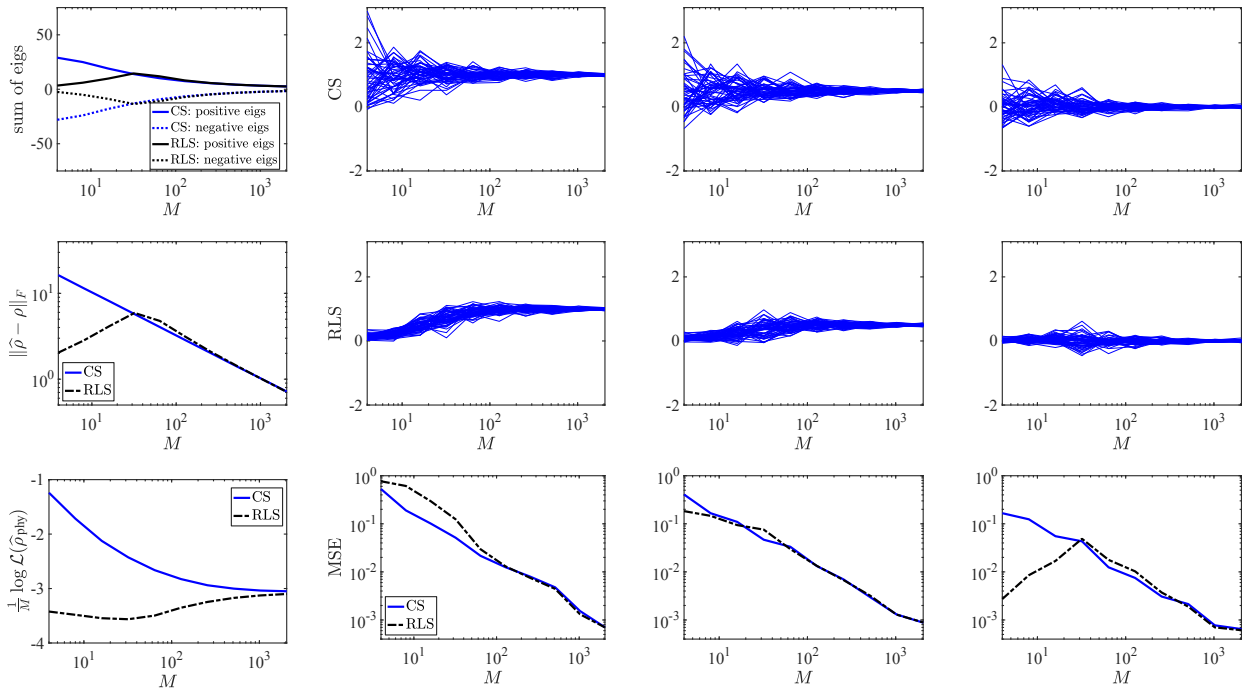
Feature 1: bias and variance tradeoff.—CS and RLS approaches trade off bias and variance in opposite ways. CS estimates are always unbiased but can exhibit relatively large variance with a limited number of measurements. On the other hand, RLS estimation controls variance through ℓ_2 regularization, but also introduces bias.

Feature 2: handling of distribution mismatch.—The quantum channel $\mathcal{M}$ in CS relies on information about how the POVMs are randomly generated. Such information is not necessary in RLS. In other words, RLS is more flexible as it only requires the POVMs actually used, regardless of whether they are generated randomly or deterministically. On the other hand, by averaging over all possible POVMs, the quantum channel in CS often has a simple explicit formulation that is independent of the specific POVMs measured, as shown in Eqs. (17, 18). In contrast, RLS must compute the inverse $(\mathcal{A}^\dagger\mathcal{A} + \mu\mathbf{I})^{-1}$ for each POVM realization. Thus, RLS and CS trade off flexibility in the distribution with computational efficiency.

Feature 3: scaling with multishot measurements.—The measurement of M POVMs with L shots each requires a total of ML state preparations. Although we focus primarily on the $L = 1$ case—in line with the original CS formulation [15]—our derivations are completely generic with respect to L. In our third set of tests, we therefore examine the performance of both RLS and CS for multishot measurements ($L > 1$). Although *a priori* unclear to us whether RLS and CS shadows would show any differences in their respective dependencies on L, numerical tests in Sec. 4.4 find that estimation errors in RLS are much more sensitive to L—both for better and worse—than their CS counterparts.

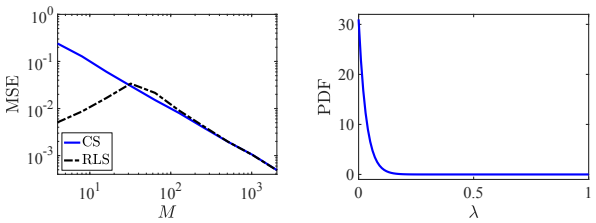
4.2 Feature 1: Bias-Variance Tradeoff

Considering the same ground truth state, observables, and measurements as analyzed by LS shadows in Fig. 2, we apply RLS and CS techniques for inference and show the results in Fig. 4. The top of the first column plots the estimates of the states in terms of their eigenvalues; the middle shows the distance to the ground truth defined as $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F$; and the bottom depicts the log-likelihood defined as $\frac{1}{M}\log\mathcal{L}(\widehat{\boldsymbol{\rho}}_{\mathrm{phy}}) =$



(a) $\widehat{\boldsymbol{\rho}}$ (b) $\widehat{\lambda}_0$ (c) $\widehat{\lambda}_1$ (d) $\widehat{\lambda}_2$

Figure 4: Illustration of the performance of CS and RLS for estimating the state $\boldsymbol{\rho}$ and the linear observables $\lambda_i = \operatorname{tr} \boldsymbol{\Lambda}_i \boldsymbol{\rho}$ with $\boldsymbol{\Lambda}_i = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\dagger$.



(a) (b)

Figure 5: (a) Illustration of the performance of the RLS and classical shadow for estimating 50 linear observables $\lambda = \operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho}), \boldsymbol{\Lambda} = \boldsymbol{\phi}\boldsymbol{\phi}^\dagger$, where $\boldsymbol{\phi}$ is randomly and uniformly generated from the unit sphere. (b) the probability distribution (i.e., probability density function (PDF)) $P(\lambda) = (D-1)(1-\lambda)^{D-2}$ for such a random linear observable λ.

$\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} f_{m,k} \log(\operatorname{tr}(\boldsymbol{A}_{m,k}\widehat{\boldsymbol{\rho}}_{\text{phy}}))$, where $\widehat{\boldsymbol{\rho}}_{\text{phy}}$ denotes the estimator projected onto physical states by the method of Ref. [67], performed to remove negative probabilities that would otherwise make the log-likelihood complex [68]. The last three columns [Fig. 4(b–d)] plot the estimates for particular expectation value λ_i for all 50 trials obtained by both RLS and CS; the bottom row shows the MSE with respect to the ground truth, averaged over all trials.

Figure 4(a) compares features of the estimators themselves. While both RLS and CS yield negative eigenvalues, those from RLS are less extreme than CS in the underdetermined regime ($M < D$) before aligning closely for $M > D$ (top). Such behavior is similarly reflected in RLS's much lower error with respect to the ground truth for $M < D$, followed by close agreement for $M > D$ (middle). Interestingly, however, as the bottom plot shows, likelihood tests (typical in classical inference) strongly favor CS over RLS; for any number of measurements M, the likelihood evaluated at the CS estimator exceeds that of RLS, despite the fact RLS is *closer* to the ground truth in much of this regime (smaller error $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_F$).

This unexpected characteristic can be explained by the fact that CS shadows are constructed directly by projectors of the measured outcomes [cf. Eq. (19)]. Hence, even for $M \ll D$, the CS estimator automatically assigns high probabilities to prior observations—which is precisely what the likelihood computes. Accordingly, the wide deviation between expectations from a likelihood test (bottom) and the actual ground truth (middle) not only reveals an interesting feature of CS shadows; it also emphasizes the importance of testing these methods against ground truth values to avoid misleading conclu-

sions on their relative merits.

On the observable side, Fig. 4(b–d) reveals that CS shadows possess large variance for low M, but are unbiased and converge to ground truth values rapidly, with nearly identical rates for all observables. This enables the derivation of rigorous information-theoretic bounds for any fixed and finite number of linear observables [15]. In contrast, the ℓ_2 regularization applied to RLS leads to observable estimators that are biased towards the origin but also have small variance. Moreover, compared to the LS shadow tests in Fig. 2, the double-decent phenomenon and wide fluctuations around $M \approx D$ have been significantly attenuated through regularization, especially striking for the λ_0 and λ_1 tests. While the RLS estimation errors for $\boldsymbol{\rho}$ and λ_2 still increase as M approaches the interpolating regime, they remain small, and they are smaller than those achieved by CS. As discussed at the end of Sec. 3.1, this increasing phase could have been mitigated by a large regularization parameter (e.g., $\mu = 1$), but at the expense of higher errors for λ_0 and λ_1.

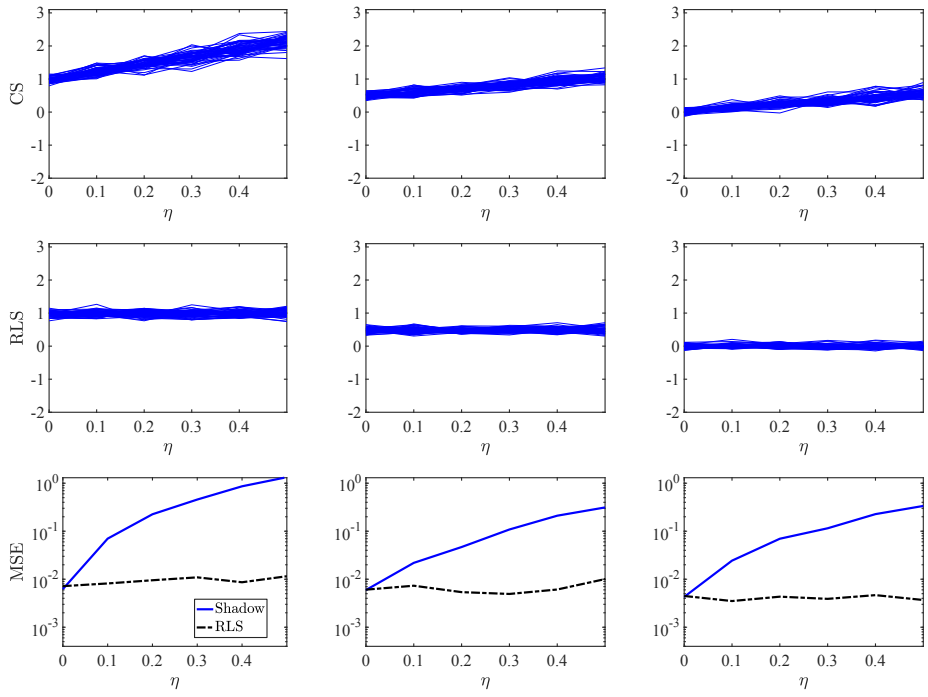
Aside: random observables.—For the numerical experiments in this paper, we have focused on a fixed state $\boldsymbol{\rho} = \boldsymbol{e}_0\boldsymbol{e}_0^\dagger$ and three rank-1, trace-1 observables with ground truth values $(\lambda_0, \lambda_1, \lambda_3) = (1, 1/2, 0)$. As discussed in detail in Ref. [19], observables with ground truth values $\lambda \sim \mathcal{O}(1)$ reflect scenarios where the accuracy of CS significantly surpasses alternative techniques like maximum likelihood and Bayesian inference for the same number of measurements. Nonetheless, from the perspective of *random* observables or ground truth states, $\lambda \approx 1$ is extremely rare in large Hilbert spaces; in this regime, for example, the Bayesian mean is far more accurate on average than CS [19].

This distinction helps explain RLS's lower MSE compared to CS for $M \ll D$ in Fig. 4(d). Taking the same simulated trials but selecting 50 different projectors of the form $\boldsymbol{\Lambda} = \boldsymbol{\phi}\boldsymbol{\phi}^\dagger$, where each $\boldsymbol{\phi}$ is randomly generated from the unit hypersphere according to the Haar measure (equivalent to looking at 50 random ground truth states for a fixed observable [19]), we find the MSEs plotted in Fig. 5(a). RLS provides more accurate estimates than CS on average when M is small and demonstrates similar performance as M becomes large. Notably, the MSEs look similar to those for λ_2 in Fig. 4(d). This can be explained by the probability density function (PDF) for the random observable $\lambda = \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho}) = \|\boldsymbol{\phi}^\dagger \boldsymbol{e}_0\|_2^2$, given by $P(\lambda) = (D-1)(1-\lambda)^{D-2}$ [69]. As depicted in Fig. 5(b), the random variable λ has a mean of $1/D = 1/32$ and mode of 0. Thus, in the random context $\lambda_2 = 0$ represents a much more typical value than either $\lambda_0 = 1$ or $\lambda_1 = 1/2$, justifying the strikingly similar behavior for $\widehat{\lambda}_2$ [Fig. 4(d)] compared to Haar-random cases [Fig. 5(a)]. So while we do consider observables like λ_0 which reflect situations of interest in practice (e.g., verification of high-fidelity state preparation), it is important to bear in mind the extreme improbability of this situation for truly random states or observables.

4.3 Feature 2: Distribution Mismatch

As described in Section 3.2, CS leverages the assumption of random measurements and involves the computation of the quantum channel for the entire distribution, as in Eq. (14). However, in practical experiments, the randomness is typically synthetic, and thus the measurement basis may not be perfectly generated according to the desired distribution. In this case, the distribution shift may pose robustness challenges for CS as it crucially depends on the formulation of the quantum channel. In contrast, the RLS approach does not incorporate assumptions of randomness at any point and may therefore prove more useful in this case.

To further illustrate this point, we invoke the same setup as in the previous experiments, but now generating the unitary measurement matrices from a mixture distribution $(1-\eta)P(\mathbb{U}_D) + \eta P((\mathbb{U}_2)^{\otimes n})$, where $\mathbb{U}_D$ and $(\mathbb{U}_2)^{\otimes n}$ denote global and local (tensor products of qubit) Haar-distributed unitary matrices, respectively. In other words, for each state copy we perform either a random global basis measurement with probability $1-\eta$ or a random local basis measurement with probability η. While this toy example is not expected to reflect distribution errors in practical systems, it provides a clear showcase of the key points. Suppose that an experimenter expects the chosen operations to be sampled from a global Haar distribution, but with probability η actually generates tensor products of local qubit unitaries. In this case, the CS shadows will be computed according to Eq. (19) under assumptions violated by the experiment.



(a) $\widehat{\lambda}_0$ (b) $\widehat{\lambda}_1$ (c) $\widehat{\lambda}_2$

Figure 6: Illustration of the performance of the RLS and CS approaches for estimating the linear observables $\lambda_i = \mathrm{tr}(\mathbf{\Lambda}_i \boldsymbol{\rho})$ in the presence of distribution shifts: the unitary matrices are generated from the distribution $(1-\eta)P(\mathbb{U}_D) + \eta P((\mathbb{U}_2)^{\otimes n})$, where $\mathbb{U}_D$ and $(\mathbb{U}_2)^{\otimes n}$ denote the global and local Haar-distributed unitary matrices, respectively.

Figure 6 shows numerical results for the system of interest for $M = 256$ measurements and $\eta \in (0, 0.5)$. The performance of CS degrades as a result of the distribution shift, and it worsens as η increases. In contrast, RLS depends solely on the actual measurements performed and does not rely on information about the sampling distribution, making its performance stable against such distribution shifts.

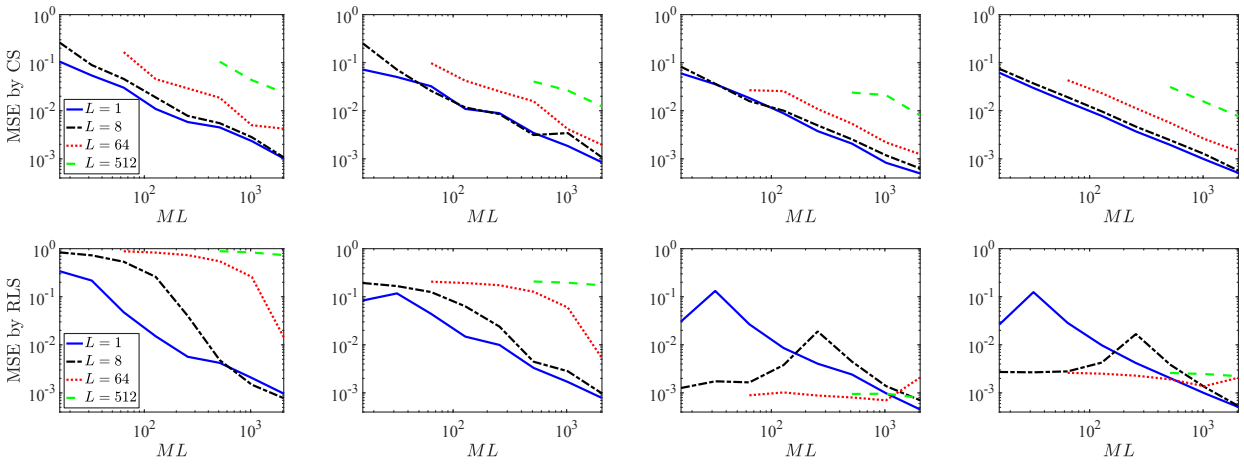
4.4 Feature 3: Multishot Measurements

The derivations culminating in Eqs. (13, 15) for RLS and CS are not confined to single-shot measurements ($L = 1$) but inherently include multishot ($L > 1$) scenarios as well. In the context of multishot measurements, where empirical frequencies are computed by averaging across all outcomes [Eq. (2)], the shadows in Eqs. (13, 15) could always be *viewed* as single-shot results but duplicated POVMs, by converting L and M to effective values $L_{\text{eff}} = 1$ and $M_{\text{eff}} = ML$. This equivalence stems from the linear nature of shadows concerning the empirical frequencies.

What then distinguishes measuring the quantum state using each POVM only once or multiple times? The primary differences are practical in origin, depending on the relative difficulty of preparing state copies compared to reconfiguring the measurement. For example, in many photonics experiments (particularly with spontaneous parametric downconversion [70, 71]), states are prepared continuously and at random, so one need only increase the integration time to push to large L for a fixed measurement setting. On the other hand, for systems composed of superconducting circuits where each state copy is actively prepared, the difference in difficulty between increasing L and increasing M is less dramatic, since both state and measurement circuit are prepared actively and deterministically.

For fixed number of state copies ML, we expect the single-shot regime $L = 1$ to provide the closest agreement between LS and CS, for in that case (maximum M) the experimental operator $\mathcal{A}^\dagger\mathcal{A}$ should approach its expected value $\mathbb{E}[\mathcal{A}^\dagger\mathcal{A}]$ most rapidly. In general, for a fixed ML, increasing M explores the Hilbert space more efficiently at the expense of greater statistical noise per setting, whereas increasing L reduces statistical noise at the expense of measurement variety.

To explore this tradeoff for CS and RLS, we



(a) $\widehat{\lambda}_0$ (b) $\widehat{\lambda}_1$ (c) $\widehat{\lambda}_2$ (d) $\widehat{\lambda}$

Figure 7: Illustration the performance of CS (top row) and RLS (bottom row) with multishot measurements for estimating the three linear observables $\{\lambda_0, \lambda_1, \lambda_2\}$ as in Fig. 2, and 50 random linear observables λ as in Fig. 5(a).

conduct numerical experiments using the same setup as in Fig. 4 in the multishot regime. In each experiment, we keep the total number of copies of the state ML fixed and vary the number of shots L within the set $\{1, 8, 64, 512\}$ for measuring each POVM, which in turn varies the number of POVMs M. Fig. 7 illustrates the performance of CS (top row) and RLS (bottom row) in estimating the expectations of the three linear observables $\{\lambda_0, \lambda_1, \lambda_2\}$ as in Fig. 2, and 50 random linear observables as in Fig. 5(a). CS consistently achieves its best performance with single-shot measurements, and its error increases with L. This observation aligns with our earlier discussion: a greater number of random POVMs brings $\frac{1}{M}\mathcal{A}^\dagger\mathcal{A}$ closer to its expected value as described in Eq. (14), and the maximum diversity of POVMs is attained in a single-shot measurement.

Interestingly, while the primary impact of $L > 1$ for CS estimation is to shift the total error up for a given ML, the log-log slopes of all examples remain approximately -1, indicating favorable scaling MSE $\propto (ML)^{-1}$ for all L regimes considered. On the other hand, $L > 1$ examples alter the *slopes* of the RLS estimator errors as well as their absolute values. For the $\widehat{\lambda}_0$ and $\widehat{\lambda}_1$ cases, the error increases rapidly with L accompanied by an extremely shallow initial slope [Fig. 7(a,b)]; in contrast, the low-ML regime of MSE for $\widehat{\lambda}_2$ and random linear observables $\widehat{\lambda}$ is actually *lower* for $L > 1$ compared to $L = 1$ [Fig. 7(c)]. Both of these features are likely due to the bias present in RLS shadows. With a smaller number of POVMs (low M), the observables computed by RLS are heavily biased to zero, which happens to deviate strongly from the ground truth values $\lambda_0 = 1$ and $\lambda_1 = 1/2$, yet is precisely the true expectation of $\mathbf{\Lambda}_2$ ($\lambda_2 = 0$) and close to the expectation of most random $\mathbf{\Lambda}$ [cf. Fig. 5(b)]. In contrast, because the CS shadow is always unbiased regardless of M and L, comparable behavior is seen for all examined observables.

To complement these numerical and qualitative findings, we now mathematically derive MSE formulas for the expectation of observable $\mathbf{\Lambda}$ [$\lambda = \mathrm{tr}(\mathbf{\Lambda}\boldsymbol{\rho})$] with estimator $\widehat{\lambda} = \mathrm{tr}(\mathbf{\Lambda}\widehat{\boldsymbol{\rho}}) = \frac{1}{M}\sum_{m=1}^{M}\mathrm{tr}(\mathbf{\Lambda}\widehat{\boldsymbol{\rho}}_m)$. We will focus on CS shadows as they are unbiased estimators, which will simplify the analysis, and exhibit consistent behavior across different observables in Fig. 7. The following result establishes the variance (and hence MSE as CS is unbiased) of CS for estimating λ.

Theorem 1. *Consider a ground truth state $\boldsymbol{\rho}$ which is repeatedly prepared and measured with M POVMs $\{\boldsymbol{A}_{m,k}\}_{k\in[K]}, m \in [M]$ generated independently and randomly from an ensemble $\mathbb{A}$ according to probability distribution $P(\mathbb{A})$. Each POVM is used to measure the state L times. Then the MSE of the CS estimate of the expectation of the linear observable $\mathbf{\Lambda}$ is given by*

$$
\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\operatorname{tr}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\rho}})-\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})\right)^2\right]\\
&=\frac{1}{ML}\,\mathbb{E}_{\{\boldsymbol{A}_k\}\sim P(\mathbb{A})}\left[\sum_{k=1}^{K}\left(\operatorname{tr}(\boldsymbol{A}_k\boldsymbol{\rho})+(L-1)\left(\operatorname{tr}(\boldsymbol{A}_k\boldsymbol{\rho})\right)^2\right)\cdot\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\right)^2\right]\\
&\quad+\frac{1-1/L}{M}\,\mathbb{E}_{\{\boldsymbol{A}_k\}\sim P(\mathbb{A})}\left[\sum_{k\neq k'}\operatorname{tr}(\boldsymbol{A}_k\boldsymbol{\rho})\cdot\operatorname{tr}(\boldsymbol{A}_{k'}\boldsymbol{\rho})\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_{k'}\right)\right)\right]\\
&\quad-\frac{1}{M}\left(\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})\right)^2.
\end{aligned}
\tag{20}
$$

Proof of Theorem 1. Using the expression $\widehat{\boldsymbol{\rho}}=\frac{1}{M}\sum_m\widehat{\boldsymbol{\rho}}_m$, we have

$$
\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\frac{1}{M}\sum_{m=1}^{M}\operatorname{tr}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\rho}}_m)-\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})\right)^2\right]\\
&=\frac{1}{M^2}\sum_m\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\operatorname{tr}(\boldsymbol{\Lambda}\widehat{\boldsymbol{\rho}}_m)\right)^2\right]-\frac{1}{M}\left(\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})\right)^2\\
&=\frac{1}{M^2}\sum_m\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\sum_{k=1}^{K}\widehat{p}_{m,k}\boldsymbol{A}_{m,k}\right)\right)\right)^2\right]-\frac{1}{M}\left(\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\rho})\right)^2,
\end{aligned}
\tag{21}
$$

where the first equality follows from the independence and unbiasedness of the shadows $\widehat{\boldsymbol{\rho}}_m, m\in[M]$. We now focus on the analysis of $\mathbb{E}_{\{\boldsymbol{A}_{m,k}\}\sim P(\mathbb{A}),\widehat{\boldsymbol{p}}_m}\left[\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\sum_{k=1}^{K}\widehat{p}_{m,k}\boldsymbol{A}_{m,k}\right)\right)\right)^2\right]$. Since this term will be the same for each m, for simplicity, we drop the subscript m and write it as $\mathbb{E}\left[\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\sum_{k=1}^{K}\widehat{p}_k\boldsymbol{A}_k\right)\right)\right)^2\right]$, where $\{\widehat{p}_k\}$ are the empirical frequencies that are obtained by using the randomly generated POVM $\{\boldsymbol{A}_k\}$ to measure the quantum state L times. We note that here the expectation is taken over two types of randomness: the randomly selected POVM $\{\boldsymbol{A}_k\}$ and the random measurements $\{\widehat{p}_k\}$. Conditioned on $\{\boldsymbol{A}_k\}$, $\{\widehat{p}_k\}$ obeys a multinominal distribution with properties

$$
\mathbb{E}\left[\widehat{p}_k^2\mid\{\boldsymbol{A}_k\}\right]=p_k^2+\frac{p_k(1-p_k)}{L}=\frac{p_k+(L-1)p_k^2}{L},
\tag{22}
$$

$$
\mathbb{E}\left[\widehat{p}_k\widehat{p}_{k'}\mid\{\boldsymbol{A}_k\}\right]=p_kp_{k'}-\frac{1}{L}p_kp_{k'}=\left(1-\frac{1}{L}\right)p_kp_{k'},\ \forall\ k\neq k'.
\tag{23}
$$

We now proceed by using these results and the fact that $\mathcal{M}^{-1}$ is a linear operator:

$$
\begin{aligned}
&\mathbb{E}_{\{\boldsymbol{A}_k\},\widehat{\boldsymbol{p}}}\left[\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\sum_{k=1}^{K}\widehat{p}_k\boldsymbol{A}_k\right)\right)\right)^2\right]=\mathbb{E}_{\{\boldsymbol{A}_k\},\widehat{\boldsymbol{p}}}\left[\left(\sum_{k=1}^{K}\widehat{p}_k\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\right)^2\right]\\
&=\mathbb{E}_{\{\boldsymbol{A}_k\},\widehat{\boldsymbol{p}}}\left[\sum_{k=1}^{K}\widehat{p}_k^2\cdot\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\right)^2\right]\\
&\quad+\mathbb{E}_{\{\boldsymbol{A}_k\},\widehat{\boldsymbol{p}}}\left[\sum_{k\neq k'}\widehat{p}_k\widehat{p}_{k'}\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_{k'}\right)\right)\right]\\
&=\mathbb{E}_{\{\boldsymbol{A}_k\}}\left[\sum_{k=1}^{K}\frac{p_k+(L-1)p_k^2}{L}\cdot\left(\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\right)^2\right]\\
&\quad+\mathbb{E}_{\{\boldsymbol{A}_k\}}\left[\sum_{k\neq k'}\left(1-\frac{1}{L}\right)p_kp_{k'}\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\cdot\operatorname{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_{k'}\right)\right)\right].
\end{aligned}
\tag{24}
$$

We complete the proof by plugging the above into Eq. (21). □

When $L = 1$, Eq. (20) reduces to the formulation in Ref. [15] (Lemma S1 therein) through the property that $\mathcal{M}^{-1}$ is self-adjoint and the expression for rank-1 orthonormal POVMs [Eq. (19)]. During the final preparation and revision of this work, we became aware of two works with derivations similar to ours: one for rank-1 orthonormal POVMs with multishot measurements [24] and another that focuses on MSE for a single unitary and then studies how the MSE varies with different choices of unitary groups [72]. Our formulation in Eq. (20) is distinct by holding for general POVMs. Though Eq. (20) may appear complex, if we disregard the last term, we can draw the following two key observations. (i) Since $\mathrm{tr}(\boldsymbol{A}_k\boldsymbol{\rho})$ is often very small, the dominant term becomes $\frac{1}{ML}\,\mathbb{E}_{\boldsymbol{A}_k}\left[\sum_{k=1}^{K}\mathrm{tr}(\boldsymbol{A}_k\boldsymbol{\rho})\cdot\left(\mathrm{tr}\left(\boldsymbol{\Lambda}\cdot\mathcal{M}^{-1}\left(\boldsymbol{A}_k\right)\right)\right)^2\right]$. This term decreases proportionally to $1/ML$, where ML represents the total number of measurements. (ii) Conversely, if we keep ML fixed, using a larger value of L generally results in a larger MSE since the remaining terms increase with L. This explains the observed decrease in performance with increasing L as shown in Figure 7.

5 Conclusion

In this paper, we have identified and formalized deep connections between traditional LS-based techniques for quantum state estimation and the disruptive methodologies of CS. Through careful derivation of the LS tomographic problem, we have shown that the LS estimator can be viewed as the average of distinct "shadows" $\widehat{\boldsymbol{\rho}}_m$, each corresponding to a specific measurement, in complete analogy with CS. This extension of the shadow picture to LS in turn reveals a novel viewpoint for CS in connection with regularization; just like traditional techniques such as RLS, CS reduces the instabilities of LS in the underdetermined regime through replacement of $(\frac{1}{M}\mathcal{A}^\dagger\mathcal{A})^+$ with a well-conditioned channel inverse.

Notwithstanding these intuitive similarities between RLS and CS shadows, our tests above reveal key differences. RLS shadows reduce variance at the cost of bias, are robust to errors in the distribution of random measurements, and are highly sensitive to the tradeoff in the number of POVMs M and number of shots L. In contrast, CS shadows are unbiased at the expense of variance, produce high estimation errors whenever the actual measurements diverge from the expected distribution, and scale favorably with a variety of M and L combinations. Certainly, although not optimal in all categories of interest, the fact that CS shadows are unbiased for any number of measurements—even in the highly underdetermined regime—is a remarkable feature that distinguishes its version of regularization from alternatives such as RLS.

Irrespective of such observations, none of the various tradeoffs can minimize the exceptional *computational* efficiency possible with CS shadows over both LS and RLS methods. Whenever the quantum channel $\mathcal{M}$ can be analytically inverted—as in the example in Eqs. (17,18)—CS requires numerical calculation of no matrix inverses, unlike both LS [Eq. (10)] and RLS [Eq. (13)]. Indeed, while examples of CS shadows up to 120 qubits were shown in Ref. [15], the record dimensionality for LS tomography (specifically, LS projected onto physical states) is a comparatively meager 14 qubits [73], and it is difficult to imagine significant increases beyond that number with existing computing technology. However, while CS shadows may face minimal competition in ultralarge Hilbert spaces, our findings connecting it to LS methods reveal a fascinating conceptual lineage with traditional methodologies, shedding further light into the secrets of the exciting and transformative tomographic procedure that is CS.

Code Availability

The MATLAB code used to produce the results in this study is available at https://github.com/ZhihuiZhu/shadow_ls.

Acknowledgments

We acknowledge funding support from the National Science Foundation (CCF-2241298, EECS-2409701), a Partnership Seed Award from the Center for Quantum Information and Engineering (CQISE) at the Ohio State University, and the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ERKJ432, ERKJ353, DE-SC0024257). We thank the Ohio Supercomputer Center for

providing the computational resources and the Quantum Collaborative led by Arizona State University for providing valuable expertise and resources. A portion of this work was performed at Oak Ridge National Laboratory, operated by UT-Battelle for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We are grateful to Stephen Becker, Zhexuan Gong, Zhen Qin, Michael Wakin, Otfried Gühne and Nikolai Wyderka for many valuable discussions.

References

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge, 2000).

[2] S. Aaronson, in *Proc. 50th Ann. ACM SIGACT Sym. Theor. Comput.* (ACM, 2018) pp. 325–338.

[3] P. Horodecki and A. Ekert, Phys. Rev. Lett. **89**, 127902 (2002).

[4] P. Horodecki, Phys. Rev. Lett. **90**, 167901 (2003).

[5] F. Mintert and A. Buchleitner, Phys. Rev. Lett. **98**, 140505 (2007).

[6] S. J. van Enk and C. W. Beenakker, Phys. Rev. Lett. **108**, 110503 (2012).

[7] N. Wyderka, A. Ketterer, S. Imai, J. L. Bönsel, D. E. Jones, B. T. Kirby, X.-D. Yu, and O. Gühne, Phys. Rev. Lett. **131**, 090201 (2023).

[8] A. Ketterer, N. Wyderka, and O. Gühne, Phys. Rev. Lett. **122**, 120505 (2019).

[9] S. Imai, N. Wyderka, A. Ketterer, and O. Gühne, Phys. Rev. Lett. **126**, 150501 (2021).

[10] A. Ketterer, S. Imai, N. Wyderka, and O. Gühne, Phys. Rev. A **106**, L010402 (2022).

[11] Y.-C. Liang, N. Harrigan, S. D. Bartlett, and T. Rudolph, Phys. Rev. Lett. **104**, 050401 (2010).

[12] M. C. Tran, B. Dakić, F. Arnault, W. Laskowski, and T. Paterek, Phys. Rev. A **92**, 050301 (2015).

[13] A. Seshadri, M. Ringbauer, J. Spainhour, T. Monz, and S. Becker, Phys. Rev. A **110**, 012431 (2024).

[14] P. Cieśliński, S. Imai, J. Dziewior, O. Gühne, L. Knips, W. Laskowski, J. Meinecke, T. Paterek, and T. Vértesi, arXiv:2307.01251 (2023).

[15] H.-Y. Huang, R. Kueng, and J. Preskill, Nat. Phys. **16**, 1050 (2020).

[16] A. Acharya, S. Saha, and A. M. Sengupta, Phys. Rev. A **104**, 052418 (2021).

[17] H. C. Nguyen, J. L. Bönsel, J. Steinberg, and O. Gühne, Phys. Rev. Lett. **129**, 220502 (2022).

[18] G. I. Struchalin, Y. A. Zagorovskii, E. V. Kovlakov, S. S. Straupe, and S. P. Kulik, PRX Quantum **2**, 010307 (2021).

[19] J. M. Lukens, K. J. H. Law, and R. S. Bennink, npj Quantum Inf. **7**, 113 (2021).

[20] A. Elben, R. Kueng, H.-Y. Huang, R. van Bijnen, C. Kokail, M. Dalmonte, P. Calabrese, B. Kraus, J. Preskill, P. Zoller, and B. Vermersch, Phys. Rev. Lett. **125**, 200501 (2020).

[21] T. Zhang, J. Sun, X.-X. Fang, X.-M. Zhang, X. Yuan, and H. Lu, Phys. Rev. Lett. **127**, 200501 (2021).

[22] R. Stricker, M. Meth, L. Postler, C. Edmunds, C. Ferrie, R. Blatt, P. Schindler, T. Monz, R. Kueng, and M. Ringbauer, PRX Quantum **3**, 040310 (2022).

[23] D. Zhu, Z. P. Cian, C. Noel, A. Risinger, D. Biswas, L. Egan, Y. Zhu, A. M. Green, C. H. Alderete, N. H. Nguyen, Q. Wang, A. Maksymov, Y. Nam, M. Cetina, N. M. Linke, M. Hafezi, and C. Monroe, Nat. Commun. **13**, 6620 (2022).

[24] Y. Zhou and Q. Liu, Quantum **7**, 1044 (2023).

[25] H.-Y. Huang, R. Kueng, and J. Preskill, Phys. Rev. Lett. **127**, 030503 (2021).

[26] A. Elben, S. T. Flammia, H.-Y. Huang, R. Kueng, J. Preskill, B. Vermersch, and P. Zoller, Nat. Rev. Phys. **5**, 9 (2023).

[27] A. E. Albert, *Regression and the Moore-Penrose Pseudoinverse* (Academic Press, 1972).

[28] C. Schwemmer, L. Knips, D. Richart, H. Weinfurter, T. Moroder, M. Kleinmann, and O. Gühne, Phys. Rev. Lett. **114**, 080403 (2015).

[29] H. Akaike, IEEE Trans. Auto. Contr. **19**, 716 (1974).

[30] J. O. S. Yin and S. J. van Enk, Phys. Rev. A **83**, 062110 (2011).

[31] S. J. van Enk and R. Blume-Kohout, New J. Phys. **15**, 025024 (2013).
[32] T. L. Scholten and R. Blume-Kohout, New J. Phys. **20**, 023050 (2018).
[33] H. Yano and N. Yamamoto, J. Phys. A: Math. Theor. **56**, 405301 (2023).
[34] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Proc. Natl. Acad. Sci. U.S.A. **116**, 15849 (2019).
[35] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, J. Stat. Mech. **2021**, 124003 (2021).
[36] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma, arXiv:2002.11328 (2020).
[37] S. P. Singh, A. Lucchi, T. Hofmann, and B. Schölkopf, arXiv:2203.07337 (2022).
[38] H. Chen, Y. Bu, and G. W. Wornell, arXiv:2306.05583 (2023).
[39] P. Nakkiran, arXiv:1912.07242 (2019).
[40] M. Belkin, D. Hsu, and J. Xu, SIAM J. Math. Data Sci. **2**, 1167 (2020).
[41] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, Proc. Natl. Acad. Sci. U.S.A. **117**, 30063 (2020).
[42] R. Sonthalia, X. Li, and B. Gu, arXiv:2305.14689 (2024).
[43] A. Curth, A. Jeffares, and M. van der Schaar, arXiv:2310.18988 (2023).
[44] A. W. R. Smith, J. Gray, and M. S. Kim, PRX Quantum **2**, 020348 (2021).
[45] D. Pierangeli and C. Conti, Nat. Commun. **14**, 1831 (2023).
[46] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, Phys. Rev. Lett. **105**, 150401 (2010).
[47] Y.-K. Liu, arXiv:1103.2816 (2011).
[48] R. Kueng, H. Rauhut, and U. Terstiege, Appl. Comput. Harmon. Anal. **41**, 88 (2017).
[49] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu, IEEE Trans. Inf. Theory **63**, 5628 (2017).
[50] M. Guţă, J. Kahn, R. Kueng, and J. A. Tropp, J. Phys. A: Math. Theor. **53**, 204001 (2020).
[51] D. S. França, F. G. Brandão, and R. Kueng, arXiv:2009.08216 (2021).
[52] F. Verstraete, J. J. Garcia-Ripoll, and J. I. Cirac, Phys. Rev. Lett. **93**, 207204 (2004).
[53] B. Pirvu, V. Murg, J. I. Cirac, and F. Verstraete, New J. Phys. **12**, 025012 (2010).
[54] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, Nat. Commun. **1**, 149 (2010).
[55] T. Baumgratz, D. Gross, M. Cramer, and M. B. Plenio, Phys. Rev. Lett. **111**, 020401 (2013).
[56] A. Lidiak, C. Jameson, Z. Qin, G. Tang, M. B. Wakin, Z. Zhu, and Z. Gong, arXiv:2207.06397 (2022).
[57] K. Noh, L. Jiang, and B. Fefferman, Quantum **4**, 318 (2020).
[58] J. Wang, Z.-Y. Han, S.-B. Wang, Z. Li, L.-Z. Mu, H. Fan, and L. Wang, Phys. Rev. A **101**, 032321 (2020).
[59] J. G. Jarkovskỳ, A. Molnár, N. Schuch, and J. I. Cirac, PRX Quantum **1**, 010304 (2020).
[60] Z. Qin, C. Jameson, Z. Gong, M. B. Wakin, and Z. Zhu, arXiv:2306.09432 (2024).
[61] T. Opatrnỳ, D.-G. Welsch, and W. Vogel, Phys. Rev. A **56**, 1788 (1997).
[62] B. Mu, H. Qi, I. R. Petersen, and G. Shi, Automatica **114**, 108837 (2020).
[63] P. Nakkiran, P. Venkat, S. Kakade, and T. Ma, arXiv:2003.01897 (2021).
[64] A. Zhao, N. C. Rubin, and A. Miyake, Phys. Rev. Lett. **127**, 110504 (2021).
[65] H.-Y. Hu, S. Choi, and Y.-Z. You, Phys. Rev. Research **5**, 023027 (2023).
[66] K. Bu, D. E. Koh, R. J. Garcia, and A. Jaffe, npj Quantum Inf. **10**, 6 (2024).
[67] J. A. Smolin, J. M. Gambetta, and G. Smith, Phys. Rev. Lett. **108**, 070502 (2012).
[68] Other approaches to obtain a meaningful likelihood from a nonphysical estimate are possible, such as removing terms with negative probabilities. However, we have found this approach to lead to the same general conclusions and so focus on physical projection only in the main text.
[69] K. Życzkowski and H.-J. Sommers, Phys. Rev. A **71**, 032313 (2005).
[70] L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, Cambridge, UK, 1995).
[71] Y. Shih, Rep. Prog. Phys. **66**, 1009 (2003).
[72] J. Helsen and M. Walter, Phys. Rev. Lett. **131**, 240602 (2023).
[73] Z. Hou, H.-S. Zhong, Y. Tian, D. Dong, B. Qi, L. Li, Y. Wang, F. Nori, G.-Y. Xi-

ang, C.-F. Li, and G.-C. Guo, New J. Phys. **18**, 083036 (2016).