Stable Tomography for Structured Quantum States

Zhen Qin, Casey Jameson, Zhexuan Gong, Michael B. Wakin and Zhihui Zhu*

Abstract

The reconstruction of quantum states from experimental measurements, often achieved using quantum state tomography (QST), is crucial for the verification and benchmarking of quantum devices. However, performing QST for a generic unstructured quantum state requires an enormous number of state copies that grows *exponentially* with the number of individual quanta in the system, even for the most optimal measurement settings. Fortunately, many physical quantum states, such as states generated by noisy, intermediate-scale quantum computers, are usually structured. In one dimension, such states are expected to be well approximated by matrix product operators (MPOs) with a finite matrix/bond dimension independent of the number of qubits, therefore enabling efficient state representation. Nevertheless, it is still unclear whether efficient QST can be performed for these states in general. In other words, there exist no rigorous bounds on the number of state copies required for reconstructing MPO states that scales polynomially with the number of qubits.

In this paper, we attempt to bridge this gap and establish theoretical guarantees for the stable recovery of MPOs using tools from compressive sensing and the theory of empirical processes. We begin by studying two types of random measurement settings: Gaussian measurements and Haar random rank-one Positive Operator Valued Measures (POVMs). We show that the information contained in an MPO with a finite bond dimension can be preserved using a number of random measurements that depends only *linearly* on the number of qubits, assuming no statistical error of the measurements. We then study MPO-based QST with physical quantum measurements through Haar random rank-one POVMs that can be implemented on quantum computers. We prove that only a *polynomial* number of state copies in the number of qubits is required to guarantee bounded recovery error of an MPO state. Remarkably, such recovery can be achieved by performing each random POVM only once, despite the large statistical error associated with the outcome of each measurement. Our work may be generalized to accommodate random local or t-design measurements that are more practical to implement on current quantum computers. It may also facilitate the discovery of efficient QST methods for other structured quantum states.

1 Introduction

Driven by advances in hardware and experimental techniques, the size of quantum computers has rapidly increased in recent years, with some of the most advanced processors having over 100 qubits [1–3]. As quantum computing and quantum simulation continue to advance, fully characterizing the large quantum many-body states produced by experimental quantum devices has become a significant challenge, as the number of parameters needed to characterize these states scales exponentially in the number of qubits in general. Nevertheless, for verification and benchmarking purposes, it is important to reconstruct such quantum states with an affordable amount of resources and with high accuracy.

The reconstruction of quantum states is typically achieved by a technique known as quantum state tomography (QST) [4]. A standard QST problem aims to find a density matrix that describes the quantum state under interest with high accuracy.[1] In a quantum system consisting of n qudits (which are d-level quantum systems; qubits have $d = 2$), the state can be expressed by a density matrix $\boldsymbol{\rho}$ of size $d^n \times d^n$. To find $\boldsymbol{\rho}$ of an experimental quantum state, in general we need to perform quantum measurements on many identical copies of the state. Any physical measurement on a quantum system is described by a Positive Operator-Valued Measure (POVM), which is a collection of positive semi-definite (PSD) matrices or operators $\{\boldsymbol{A}_1, \dots, \boldsymbol{A}_K\}$ that sum to the identity operator. Each operator

*ZQ (email: qin.660@osu.edu) and ZZ (email: zhu.3440@osu.edu) are with the Department of Computer Science and Engineering, the Ohio State University; CJ (email: cwjameson@mines.edu) and ZG (email: gong@mines.edu) are with the Department of Physics, Colorado School of Mines; and MBW (email: mwakin@mines.edu) is with the Department of Electrical Engineering, Colorado School of Mines.

[1] See Section 2 for an overview of basic quantum mechanics.

$\boldsymbol{A}_k$ $(k = 1, \dots, K)$ in the POVM corresponds to a possible measurement outcome, and the probability of obtaining that outcome is given by $p_k = \mathrm{trace}(\boldsymbol{A}_k\boldsymbol{\rho})$. Thus, this *probabilistic nature* of quantum measurements often requires the state to be measured many (say M) times with the same POVM to obtain an approximately accurate statistical estimate $\widehat{p}_k$ of each p_k. Without considering the statistical error, $\{p_k\}$ can be viewed as K linear measurements of the state $\boldsymbol{\rho}$. Thus, adopting terminology from machine learning, we may refer to $\{p_k\}$ and their empirical estimates $\{\widehat{p}_k\}$ as population and empirical measurements of the state, respectively. From this viewpoint, QST can be viewed as a matrix sensing problem [5,6], but with a specific type of measurement operator, and with measurements that are inherently probabilistic. Furthermore, according to quantum mechanics, when a projective measurement is performed on a quantum state, the state collapses to one of the possible eigenstates of the measured observable, resulting in a different state in general. Therefore, we need many identical copies of the state for performing many measurements. Typically, an interesting quantum many-body state can be generated using a quantum computer or quantum simulator in a time scale ranging from microseconds to milliseconds for common hardware platforms. If the number of state copies required by QST scales exponentially in the number of qudits, then we cannot perform QST in practice for even a few tens of qubits.

Many different methods have been proposed for QST, including maximum likelihood [7,8], Bayesian [9–11], region [12,13], and least squares [14,15] estimators, and machine learning techniques [16–18]. For generic quantum states, the number of state copies needed for QST always grows exponentially with the number of qudits. A significant amount of work has been dedicated, however, to optimal QST methods for states represented by low-rank density matrices, which are physically common [19–23]. Various measurement settings have been adopted in this context, including 4-design [19], Pauli [20,24], Clifford [21], Haar-distributed unitary [22], etc. It has been shown that as long as the measurements are performed on one state at a time, a minimum number of total state copies proportional to $d^n r^2/\epsilon^2$ is required to estimate a rank-r density matrix with accuracy given by ϵ in the trace norm between the reconstructed density matrix and the true density matrix [21,23]. This means that even for a rank-one density matrix (corresponding to a pure quantum state that can only be created by a noiseless quantum device), the number of state copies required for QST still scales as 2^n for n qubits.

To achieve QST for current quantum computers at the scale of $\sim$100 qubits, the number of required state copies should scale only polynomially with the number of qubits n. This is possible only if the target state itself is structured in a way such that it has a compact representation with $\mathrm{poly}(n)$ independent parameters. Fortunately, many physical quantum states indeed have such structure. Examples include ground states of most quantum systems with short-range interactions and states generated by such quantum systems in a finite amount of time [25]. These states usually do not contain a large amount of quantum entanglement such that a compact representation via a matrix product state (MPS) or tensor network is often possible [25]. A similar intuition applies to states generated by noisy quantum computers, where the noise could also limit the amount of quantum entanglement and thus enable an efficient state representation. In particular, it has been recently shown that states generated by a one-dimensional noisy quantum computer are well approximated by matrix product operators (MPOs) with a finite matrix dimension [26]. Therefore, it becomes practically important to find efficient QST methods for states with an efficient MPO representation.

An MPO consists of nd^2 matrices each with dimension at most $\overline{r} \times \overline{r}$. The matrix dimension $\overline{r}$ is more often called the bond dimension, or the rank of the MPO (see Section 2.3 for the detailed description of MPO). The MPO is also mathematically equivalent to the tensor train (TT) used for compact representation of large tensors [27]. Assuming the bond dimension $\overline{r}$ is finite, the MPO contains a number of parameters that scales only linearly with the number of qudits, and is thus a very efficient representation. Nevertheless, such an efficient representation does not guarantee that the number of state copies required for QST is also small. In fact, for a general MPO state with bond dimension $\overline{r}$, there exists no known QST method that guarantees a required number of state copies that scales polynomially with the number of qubits [28,29]. This is in contrast to an MPS state (a pure state with a compact representation using nd matrices), where such a guarantee exists for almost all physical MPS states [30–36]. Therefore, we ask the following main question:

Question: Given a structured n-qudit quantum state represented by a finite bond dimension MPO, is it possible to reconstruct the state with guaranteed accuracy using only $\mathrm{poly}(n)$ state copies?

1.1 Main results

In this paper, we show that the answer to the above main question is yes, assuming that we can perform measurements of the given quantum state in Haar random bases. We note that this affirmative answer does not imply efficient QST for general MPO states since an exponentially large number (in n) of local quantum gates may be required to achieve such Haar random basis measurements. Nevertheless, our results paves the way to fully efficient QST methods as one may be able to reduce such number of required local quantum gates to polynomial in n via unitary t-designs [37].

Our particular focus on Haar random bases is motivated by the tremendous success of randomized measurements in compressive sensing for signals exhibiting low-dimensional structure such as sparse, low rank, or manifold structure [5, 38–42]. The incorporation of randomness often enables nearly optimal upper bounds to be established for the sufficient number of measurements to recover structured signals. Moreover, randomized measurements have been recognized as a powerful tool that can efficiently transform quantum systems into classical representations, capturing numerous features of the original quantum state [21,43,44]; see [45] for a review on this topic.

The first main contribution of this paper—presented in Section 3—is that we investigate the number of population measurements (without statistical errors) to guarantee a stable embedding of MPOs. In particular, we first establish the restricted isometry property (RIP, see Definition 2) for complex Gaussian measurements where each matrix element of $\boldsymbol{A}_k$ is an independent and identically distributed (*i.i.d.*) standard complex Gaussian random variable for all $k = 1, \dots, K$. Although these measurement operators are not PSD and may not be implementable in practical quantum experiments, this analysis sheds light on the optimal number of population measurements to ensure unique recovery of the MPO. We then study rank-one Gaussian measurement ensembles $\{\boldsymbol{A}_k\}$ taking the form $\boldsymbol{A}_k = \boldsymbol{a}_k \boldsymbol{a}_k^H$ where $\boldsymbol{a}_k$ is randomly generated from a multivariate Gaussian distribution. As such rank-one measurements do not obey the RIP condition [46], we instead establish a weaker version of an embedding guarantee. In order to do this, we use Mendelson's small ball method [42,47,48], which has previously been used to establish stable embeddings for low-rank matrices under rank-one measurements [19]. For both generic Gaussian measurement ensembles and rank-one Gaussian measurement ensembles, we show that $\widetilde{\Omega}(nd^2\overline{r}^2)$ total linear measurements[2] are sufficient to achieve stable embeddings of MPOs with high probability. This result is nearly optimal as the MPO contains $nd^2\overline{r}^2$ independent parameters.

We then extend the results to Haar random rank-one POVMs, where each POVM is a collection of PSD matrices $\{\boldsymbol{\phi}_k \boldsymbol{\phi}_k^H\}, k = 1, \dots, d^n$ with $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{d^n} \end{bmatrix}$ being a Haar-distributed random unitary matrix. As will be formally illustrated in Section 2, such a measurement scheme is equivalent to first rotating the state with the unitary matrix $\boldsymbol{\Phi}$ and then performing measurements in the standard computational basis, which can be implemented (albeit not efficiently) on current quantum computers [21]. We establish similar stable embedding results for $Q = \widetilde{\Omega}(nd^2\overline{r}^2)$ such random rank-one POVMs, assuming zero statistical error.

Second, we study the recovery of an MPO from empirical quantum measurements (physical measurements containing statistical errors) and establish recovery bounds with respect to the number of state copies, using the above-mentioned Haar random measurement bases. *The second main contribution of this paper—presented in Section 4—is that we establish theoretical bounds on the accuracy of a particular estimator—the solution to a constrained least-squares optimization problem—for recovering an MPO.* We summarize the results informally as follows.

Theorem 1 (informal version of Theorem 5)**.** *Given an n-qudit MPO state with bond dimension $\overline{r}$, randomly generate Q Haar random rank-one POVMs and perform measurements with each POVM M times. For any $\epsilon > 0$, assume $Q = \widetilde{\Omega}(nd^2\overline{r}^2)$ and the number of total state copies $QM = \widetilde{\Omega}(n^3d^2\overline{r}^2/\epsilon^2)$. Then, with high probability, a properly constrained least-squares minimization with the empirical measurements stably recovers the ground-truth state with ϵ-closeness in the Frobenius norm.*

Our result ensures a stable recovery of the ground-truth state with a total number of state copies QM growing only polynomially in the number of qudits n. Compared to the requirement of $\Omega(d^n)$ state copies for estimating a low-rank state, utilizing the MPO structure can significantly reduce the number of state copies (from d^n to n^3). In addition, there is no other requirement on the number of state copies M for each POVM. In other words, our result also provides theoretical support for the practical use of single-shot measurements (setting $M = 1$, i.e., measuring each POVM only once) that have been practically adopted in [32,43].

We note that obtaining the constrained least squares estimate requires solving a nonconvex problem. To tackle this problem, we employ iterative hard thresholding (i.e., projected gradient descent) [49] and showcase its efficacy through

[2] The notation $\widetilde{\Omega}(\cdot)$ is defined in Section 1.3.

numerical experiments. We do not provide a formal guarantee for the algorithm and leave its analysis for future work.

1.2 Related work involving tensor train decompositions

Having mentioned that the MPO model is equivalent to a tensor train (TT) decomposition, we discuss some related work on sampling and recovery of tensors. The work [49] established the first RIP bound for structured tensors (including the TT format) with real generic subgaussian measurements. Our proof of the RIP for complex Gaussian measurements uses the same technique as [49]; see the discussion following Theorem 6 for more information. The work [50] studied the tensor completion problem with random samples of a TT format tensor, but the result requires an exponentially large number of samples. Another line of work [51–53] extended matrix *cross approximation* techniques [54–56] for computing a TT format from selected subtensors. The work [57] has provided accuracy guarantees in terms of the entire tensor for TT cross approximation, and the work [29] applied TT cross approximation for reconstructing MPOs by only measuring local operators. Numerical simulation results demonstrate the effectiveness of this technique, but no explicit theoretical bound on the number of state copies is provided [29]. While the algorithm is not the focus of this work, we note that there are many proposed algorithms for estimating TT format tensors from linear measurements [49,50,58–63]. These include algorithms based on convex relaxation [58,59], alternating minimization [60], projected gradient descent (also known as iterative hard thresholding (IHT)) [49], and Riemannian methods [50,62,63].

1.3 Notation

We use calligraphic letters (e.g., $\mathcal{X}$) to denote tensors, bold capital letters (e.g., $\boldsymbol{X}$) to denote matrices, bold lowercase letters (e.g., $\boldsymbol{x}$) to denote vectors, and italic letters (e.g., x) to denote scalar quantities. Elements of matrices and tensors are denoted in parentheses, as in Matlab notation. For example, $\mathcal{X}(i_1, i_2, i_3)$ denotes the element in position (i_1, i_2, i_3) of the order-3 tensor $\mathcal{X}$. The calligraphic letter $\mathcal{A}$ is reserved for the linear measurement map. For a positive integer K, $[K]$ denotes the set $\{1, \dots, K\}$. The superscripts $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and Hermitian transpose operators, respectively. For two matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same size, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \operatorname{trace}(\boldsymbol{A}^H \boldsymbol{B})$ denotes the inner product between them. $\|\boldsymbol{A}\|$ (or $\|\boldsymbol{A}\|_{2\to 2}$) and $\|\boldsymbol{A}\|_F$ respectively represent the spectral norm and Frobenius norm of $\boldsymbol{A}$. For a vector $\boldsymbol{a}$ of size $N \times 1$, its l_n-norm is defined as $||\boldsymbol{a}||_n = (\sum_{m=1}^N |a_m|^n)^{\frac{1}{n}}$. For two positive quantities $a, b \in \mathbb{R}$, the inequality $b \lesssim a$ or $b = O(a)$ means $b \leq ca$ for some universal constant c; likewise, $b \gtrsim a$ or $b = \Omega(a)$ represents $b \geq ca$ for some universal constant c. We define $\widetilde{\Omega}$ as the function obtained by removing the logarithmic factors from Ω.

2 Quantum Mechanics

Quantum mechanics is a mathematical framework for the development of quantum theories [64]. While this subject may be unfamiliar to some researchers in information theory and signal processing, fortunately, most of its essential concepts can be understood using basic concepts from linear algebra and probability. In this section, we review the elements of quantum mechanics necessary for describing QST.

2.1 States and density operators

In quantum mechanics, the state of an isolated quantum system is fully described by a state vector $|\psi\rangle$ (using the Dirac notation), which represents a unit-length vector in a complex vector space known as the Hilbert space. For example, the state of the simplest quantum system, known as a *qubit*, is represented by a vector in a two-dimensional Hilbert space. One can choose two orthonormal basis vectors for this Hilbert space denoted by $|0\rangle$ and $|1\rangle$, which represent two distinct physical states of a qubit (e.g., the lowest and second-lowest energy states of an atom). An arbitrary state of the qubit can then be written as $|\psi\rangle = a|0\rangle + b|1\rangle$, where a and b are complex numbers satisfying $|a|^2 + |b|^2 = 1$, which ensures that $|\psi\rangle$ is unit length. The state vector $|\psi\rangle$ can thus be equivalently represented by a 2×1 vector

$$\boldsymbol{\psi} := \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{C}^2.$$

A *qudit* is a generalization of the idea of a qubit to a d-level system or d-dimensional Hilbert space, where each state vector can be equivalently represented by a unit-length vector in $\mathbb{C}^d$. While most quantum computers process information using qubits just as most classical computers use bits, we use qudits in this paper for a more general framework, as they are commonly used for quantum simulation and may be used for future quantum computers as well.

A quantum system can consist of multiple qudits. For such many-body systems, which are the focus of this paper, the full state space is the tensor product of the state spaces of each qudit. Specifically, for a composite system of n qudits, each state vector $\boldsymbol{\psi}$ belongs to $\mathbb{C}^{d^n}$ and has unit length.

Until now we have considered quantum systems whose state can be fully described by a state vector $\boldsymbol{\psi}$. Such a quantum system is said to be in a *pure state*. More broadly, though, a quantum system can be in one of a number of states $\boldsymbol{\psi}_i$ with respective probabilities α_i. In this case, we say the quantum system is in a *mixed state*, which may be described as $\{\alpha_i, \boldsymbol{\psi}_i\}$ where $0 \leq \alpha_i \leq 1$ are the probabilities with $\sum_i \alpha_i = 1$. A mixed state naturally arises due to the interactions (which create quantum entanglement) between the quantum system and its environment, such that the state of the system becomes indeterminate.

A quantum system in a mixed state is described by a *density operator* or *density matrix*.[3] The density operator of a pure state $\boldsymbol{\psi} \in \mathbb{C}^{d^n}$ is given by

$$\boldsymbol{\rho} = \boldsymbol{\psi}\boldsymbol{\psi}^H \in \mathbb{C}^{d^n \times d^n}.$$

For a mixed state, the density operator can be written as

$$\boldsymbol{\rho} = \sum_i \alpha_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^H \in \mathbb{C}^{d^n \times d^n}.$$

Thus, a density operator with rank equal to one corresponds to a pure state; otherwise it corresponds to a mixed state. In all cases, we have that (i) the density operator $\boldsymbol{\rho} \succeq \mathbf{0}$ is a PSD matrix, and (ii) $\operatorname{trace}(\boldsymbol{\rho}) = 1$.

2.2 POVM measurements

Quantum state tomography aims to construct or estimate the density operator $\boldsymbol{\rho}$ of a quantum system using measurements on an ensemble of identical quantum states. Many copies of the quantum state are needed due to the probabilistic nature of quantum measurements, which are described using Positive Operator Valued Measures (POVMs) [64].

Definition 1 (POVM and quantum measurements [64])**.** *A Positive Operator Valued Measure (POVM) is a set of PSD matrices* $\{\boldsymbol{A}_1, \dots, \boldsymbol{A}_K\}$ *such that*

$$\sum_{k=1}^{K} \boldsymbol{A}_k = \mathbf{I}. \tag{1}$$

Each POVM element $\boldsymbol{A}_k$ *is associated with a possible outcome of a quantum measurement, and the probability* p_k *of detecting the* k*-th outcome when measuring the density operator* $\boldsymbol{\rho}$ *is given by*

$$p_k = \langle \boldsymbol{A}_k, \boldsymbol{\rho} \rangle, \tag{2}$$

where $\sum_{k=1}^{K} p_k = 1$ *due to* (1) *and the fact that* $\operatorname{trace}(\boldsymbol{\rho}) = 1$*. We often repeat the measurement process* M *times and take the average of the statistically independent outcomes to generate the empirical probabilities*

$$\widehat{p}_k = \frac{f_k}{M},\ k \in [K] := \{1, \dots, K\}, \tag{3}$$

where f_k *denotes the number of times the* k*-th outcome is observed in the* M *experiments. For convenience, we call* $\{p_k\}$ *and* $\{\widehat{p}_k\}$ *the population and empirical (linear) measurements, respectively.*

Collectively, the random variables $f_1, \dots, f_K$ are characterized by the multinomial distribution $\operatorname{Multinomial}(M, \boldsymbol{p})$ [65] with parameters M and $\boldsymbol{p} = \begin{bmatrix} p_1 & \cdots & p_K \end{bmatrix}^\top$, where p_k is defined in (2). It follows that the empirical probability $\widehat{p}_k$ in (3) is an unbiased estimator of the probability p_k. One can bound the estimation error $|\widehat{p}_k - p_k|$ by $O(1/\sqrt{M})$

[3] Formally speaking, a density matrix is a representation of a density operator in a given choice of basis in the underlying Hilbert space. In this paper, we always choose the standard computational basis for the qudits denoted by $\{|0\rangle, |1\rangle, \cdots, |d-1\rangle\}$. Therefore, we use the two terms density matrix and density operator interchangeably.

with high probability via concentration inequalities. For example, the Dvoretzky-Kiefer-Wolfowith (DKW) theorem [66,67] ensures that the empirical probability $\widehat{p}_k$ is close to p_k for all k simultaneously when M is sufficiently large. In particular, for any $\epsilon > 0$,

$$\mathbb{P}\left(\max_k |p_k - \widehat{p}_k| \geq \epsilon\right) \leq 2e^{-\frac{1}{2}M\epsilon^2}. \tag{4}$$

Rank-one POVMs A particular type of POVM that is commonly used in practice is the rank-one POVM of the form $\{\boldsymbol{A}_k = \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H\}$ with $\sum_{k=1}^K \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H = \mathbf{I}$. In this case, the probability in (2) can be rewritten as

$$p_k = \langle \boldsymbol{A}_k, \boldsymbol{\rho}\rangle = \left\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\right\rangle = \boldsymbol{\phi}_k^H\boldsymbol{\rho}\boldsymbol{\phi}_k. \tag{5}$$

When $\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_K\end{bmatrix} \in \mathbb{C}^{d^n\times K}$ further forms an orthonormal basis, in which case $K = d^n$, we call $\{\boldsymbol{\phi}_k\boldsymbol{\phi}_k^H\}_{k=1}^K$ a *Haar random rank-one POVM*. In this case, it is revealing to write

$$p_k = \left\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\right\rangle = \boldsymbol{\phi}_k^H\boldsymbol{\rho}\boldsymbol{\phi}_k = \boldsymbol{e}_k^H\left(\boldsymbol{\Phi}^H\boldsymbol{\rho}\boldsymbol{\Phi}\right)\boldsymbol{e}_k, \tag{6}$$

where the last equation implies that the measurement is equivalent to first applying the unitary operator $\boldsymbol{\Phi}$ to the unknown state $\boldsymbol{\rho} \mapsto \boldsymbol{\Phi}^H\boldsymbol{\rho}\boldsymbol{\Phi}$ and then performing measurements in the canonical basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_{d^n}$. Both steps can be implemented on a universal quantum computer in practice, though the number of single and two-qubit quantum gates required of preparing the unitary matrix $\boldsymbol{\Phi}$ in general scale exponentially with the number of qubits [68]. Nevertheless, the use of Haar random rank-one POVMs is common in QST as it often provides the minimal number of required state copies [21].

Experimentally, QST always utilize the empirical probabilities $\{\widehat{p}_k\}$ in order to recover or estimate the unknown density operator $\boldsymbol{\rho}$. Note that in general the outcomes from a single POVM are not sufficient to recover the underlying density operator $\boldsymbol{\rho}$ since the number of measurements is much smaller than the size, $d^n \times d^n$, of $\boldsymbol{\rho}$. For example, a Haar random rank-one POVM only provides d^n linear measurements. Thus, a complete measurement scheme often consists of measuring the state using more than one POVM to generate more measurements.

Ensembles of POVMs Suppose we have Q POVMs $\{\boldsymbol{A}_{i,1}, \dots, \boldsymbol{A}_{i,K}\}$ for $i = 1, \dots, Q$; for simplicity, we assume that each POVM contains the same number of PSD matrices, although in general this may vary between POVMs. We use each POVM to measure a state M times to obtain the empirical measurements as described in Definition 1.

The experimental costs of acquiring measurements with ensembles of POVMs, including the required number of total state copies QM, remain prohibitively high for general states, making such measurements impractical for large quantum systems. Fortunately, practical quantum states exhibit certain low-dimensional structure that can be exploited for the inverse process. For example, the low-rank model has been widely used to reduce the number of measurements in QST [20,21,23]. However, for a low-rank density operator with rank r, at least $\Omega(d^n r^2)$ state copies are needed for stable recovery [21]. The required number of state copies grows *exponentially* with the number of qudits, making the low-rank model inefficient for large quantum systems. However, another compact representation, called the matrix product operator (MPO) [36], has emerged for approximating practical density matrices [26]. As formally described in the next subsection, the MPO representation is remarkably scalable as its number of parameters only grows *linearly* in terms of the number of qudits.

2.3 Matrix Product Operator (MPO)

For a density matrix $\boldsymbol{\rho} \in \mathbb{C}^{d^n\times d^n}$ corresponding to an n-qudit quantum system, we use a single multi-index $i_1\cdots i_n$ (correspondingly $j_1\cdots j_n$) to specify the indices of rows (correspondingly columns), where $i_1, \dots, i_n \in [d]$.[4] Then we say $\boldsymbol{\rho}$ is an MPO if we can express its $(i_1\cdots i_n, j_1\cdots j_n)$-element as the following matrix product [35]

$$\boldsymbol{\rho}(i_1\cdots i_n, j_1\cdots j_n) = \boldsymbol{X}_1^{i_1,j_1}\boldsymbol{X}_2^{i_2,j_2}\cdots\boldsymbol{X}_n^{i_n,j_n}, \tag{7}$$

where $\boldsymbol{X}_\ell^{i_\ell,j_\ell} \in \mathbb{C}^{r_{\ell-1}\times r_\ell}$ with $r_0 = r_n = 1$. See Figure 1 for an illustration. The dimensions $\boldsymbol{r} = (r_1, \dots, r_{n-1})$ are often called the *bond dimensions*[5] of the MPO in quantum physics, though we may also call them the *MPO ranks*.

[4] Specifically, $i_1\cdots i_n$ represents the $(i_1 + \sum_{\ell=2}^n d^{\ell-1}(i_\ell - 1))$-th row.

[5] It is also common to simply call $\overline{r} = \max\{r_1, \dots, r_{n-1}\}$ the bond dimension.

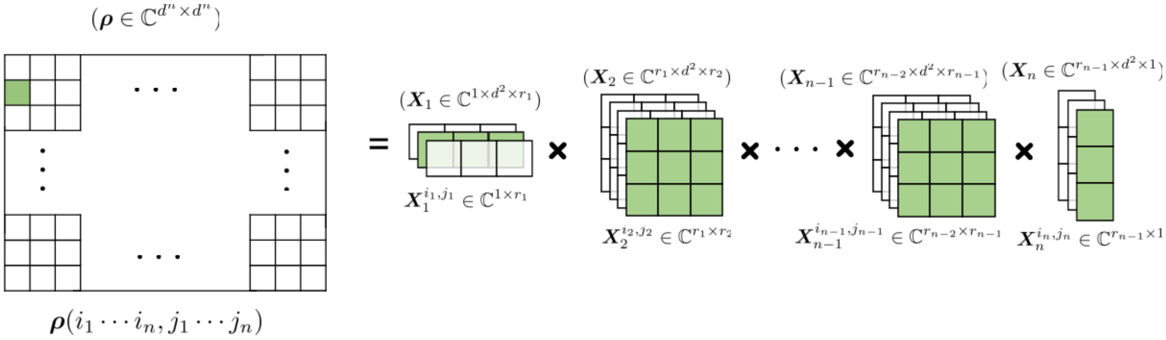

$(\boldsymbol{\rho} \in \mathbb{C}^{d^n \times d^n})$
$\boldsymbol{\rho}(i_1 \cdots i_n, j_1 \cdots j_n)$
=
$(\boldsymbol{X}_1 \in \mathbb{C}^{1 \times d^2 \times r_1})$
$\boldsymbol{X}_1^{i_1,j_1} \in \mathbb{C}^{1 \times r_1}$
×
$(\boldsymbol{X}_2 \in \mathbb{C}^{r_1 \times d^2 \times r_2})$
$\boldsymbol{X}_2^{i_2,j_2} \in \mathbb{C}^{r_1 \times r_2}$
× ⋯ ×
$(\boldsymbol{X}_{n-1} \in \mathbb{C}^{r_{n-2} \times d^2 \times r_{n-1}})$
$\boldsymbol{X}_{n-1}^{i_{n-1},j_{n-1}} \in \mathbb{C}^{r_{n-2} \times r_{n-1}}$
×
$(\boldsymbol{X}_n \in \mathbb{C}^{r_{n-1} \times d^2 \times 1})$
$\boldsymbol{X}_n^{i_n,j_n} \in \mathbb{C}^{r_{n-1} \times 1}$

Figure 1: Illustration of the MPO in (7).

These dimensions can indeed be viewed as the ranks of certain matrices that are obtained by reshaping the density matrix $\boldsymbol{\rho}$ in various ways. Rather than directly presenting the reshaped matrices, however, it is more enlightening to first discuss the connection between MPO and an equivalent form for describing tensors, the so-called *tensor train* format [27,69].

Connection to the tensor train (TT) format To illustrate the connection between the MPO and the tensor train (TT) format [27], we first reshape $\boldsymbol{\rho}$ into an n-th order tensor $\mathcal{X}$ of size $d^2 \times d^2 \times \cdots \times d^2$ by mapping each pair (i_ℓ, j_ℓ) into a single index $s_\ell = i_\ell + d(j_\ell - 1), \ell = 1, \ldots, n$ such that the elements of $\mathcal{X}$ are given by

$$\mathcal{X}(s_1, \ldots, s_n) \quad = \quad \boldsymbol{\rho}(i_1 \cdots i_n, j_1 \cdots j_n). \tag{8}$$

Note that $\mathcal{X}$ is just a reshaping of $\boldsymbol{\rho}$ and that both objects contain exactly the same entries. Then, according to (7), the $(s_1, \ldots, s_n)$-th element of $\mathcal{X}$ can also be represented as a matrix product

$$\mathcal{X}(s_1, \ldots, s_n) = \boldsymbol{X}_1^{s_1} \boldsymbol{X}_2^{s_2} \cdots \boldsymbol{X}_n^{s_n}, \tag{9}$$

where with abuse of notation we denote $\boldsymbol{X}_\ell^{s_\ell} = \boldsymbol{X}_\ell^{i_\ell, j_\ell}$. The decomposition in (9) is known as the TT decomposition and has been widely studied in the literature [27,69–73].

Canonical form When $n = 2$, the decomposition (9) is equivalent to the standard matrix factorization of the form $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{A} \in \mathbb{R}^{d^2 \times d^2}, \boldsymbol{B} \in \mathbb{R}^{d^2 \times r}, \boldsymbol{C} \in \mathbb{R}^{r \times d^2}$, the rows of $\boldsymbol{B}$ correspond to $\boldsymbol{X}_1^{s_1}$ and the columns of $\boldsymbol{C}$ correspond to $\boldsymbol{X}_2^{s_2}$. There exist infinitely many possible choices of $(\boldsymbol{B}, \boldsymbol{C})$ such that $\boldsymbol{B}\boldsymbol{C} = \boldsymbol{A}$, but all of them require $r \geq \operatorname{rank}(\boldsymbol{A})$. Among all these possible factorizations, if $\operatorname{rank}(\boldsymbol{B}) = \operatorname{rank}(\boldsymbol{C}) = r$, then $r = \operatorname{rank}(\boldsymbol{B}\boldsymbol{C}) = \operatorname{rank}(\boldsymbol{A})$, implying that this is the *minimal* r allowed for the factorization $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$. Moreover, one can always construct a factorization (say by the singular value decomposition) such that $\boldsymbol{B}$ is orthogonal with $\boldsymbol{B}^\top \boldsymbol{B} = \mathbf{I}_r$, or $\boldsymbol{C}$ is orthogonal with $\boldsymbol{C}\boldsymbol{C}^\top = \mathbf{I}_r$.

Likewise, the decomposition of the tensor $\mathcal{X}$ into the form of (9) is generally not unique: not only are the factors $\{\boldsymbol{X}_\ell^{i_\ell, j_\ell}\}$ not unique, but also the dimensions of these factors can vary. To introduce the factorization with the smallest possible dimensions $\boldsymbol{r} = (r_1, \ldots, r_{n-1})$, for convenience, for each ℓ, we put $\boldsymbol{X}_\ell = \{\boldsymbol{X}_\ell^{i_\ell, j_\ell}\}_{i_\ell, j_\ell}$ together into the following two forms

$$L(\boldsymbol{X}_\ell) = \begin{bmatrix} \boldsymbol{X}_\ell^{1,1} \\ \vdots \\ \boldsymbol{X}_\ell^{d,d} \end{bmatrix}, \; R(\boldsymbol{X}_\ell) = \begin{bmatrix} \boldsymbol{X}_\ell^{1,1} & \cdots & \boldsymbol{X}_\ell^{d,d} \end{bmatrix},$$

where $L(\boldsymbol{X}_\ell)$ and $R(\boldsymbol{X}_\ell)$ are often called the left unfolding and right unfolding of $\boldsymbol{X}_\ell$, respectively, if we view $\boldsymbol{X}_\ell$ as a tensor. We say the decomposition (9) is *minimal* if the rank of the left unfolding matrix $L(\boldsymbol{X}_\ell)$ is r_ℓ and the rank of the right unfolding matrix $R(\boldsymbol{X}_\ell)$ is $r_{\ell-1}$. The dimensions $\boldsymbol{r} = (r_1, \dots, r_{n-1})$ of such a minimal decomposition are called the *TT ranks* of $\mathcal{X}$. According to [70], there is exactly one set of ranks $\boldsymbol{r}$ that $\mathcal{X}$ admits a minimal TT decomposition. Moreover, in this case, r_ℓ equals the rank of the ℓ-th unfolding matrix $\boldsymbol{X}^{\langle \ell \rangle} \in \mathbb{C}^{d^{2\ell} \times d^{2n-2\ell}}$ of the tensor $\mathcal{X}$, where the $(s_1 \cdots s_\ell, s_{\ell+1} \cdots s_n)$-th element of $\boldsymbol{X}^{\langle \ell \rangle}$ is given by $\boldsymbol{X}^{\langle \ell \rangle}(s_1 \cdots s_\ell, s_{\ell+1} \cdots s_n) = \mathcal{X}(s_1, \dots, s_n)$. This can also serve as an alternative way to define the TT rank. As for the matrix case, for any MPO $\boldsymbol{\rho}$ of the form (7), there always exists a factorization such that $L(\boldsymbol{X}_\ell)$ are unitary matrices for all $\ell = 1, \dots, n-1$; that is

$$L(\boldsymbol{X}_\ell)^H L(\boldsymbol{X}_\ell) = \sum_{i_\ell, j_\ell} \left(\boldsymbol{X}_\ell^{i_\ell, j_\ell} \right)^H \boldsymbol{X}_\ell^{i_\ell, j_\ell} = \mathbf{I}_{r_\ell}, \ \ell = 1, \dots, n-1, \tag{10}$$

which is called the left-canonical form[6] [74]. According to [70, Theorem 1], such a canonical form is unique up to the insertion of orthogonal matrices between the factors. Thus, we will denote by $\mathbb{X}_{\overline{r}}$ the set of MPOs with maximum MPO rank equal to $\overline{r}$:

$$\begin{aligned} \mathbb{X}_{\overline{r}} = \Big\{ \boldsymbol{\rho} \in \mathbb{C}^{d^n \times d^n} : \boldsymbol{\rho} = \boldsymbol{\rho}^H, \boldsymbol{\rho}(i_1 \cdots i_n, j_1 \cdots j_n) = \boldsymbol{X}_1^{i_1, j_1} \boldsymbol{X}_2^{i_2, j_2} \cdots \boldsymbol{X}_n^{i_n, j_n}, \ \boldsymbol{X}_\ell^{i_\ell, j_\ell} \in \mathbb{C}^{r_{\ell-1} \times r_\ell}, \\ \sum_{i_\ell, j_\ell} \left(\boldsymbol{X}_\ell^{i_\ell, j_\ell} \right)^H \boldsymbol{X}_\ell^{i_\ell, j_\ell} = \mathbf{I}_{r_\ell}, \ell = 1, \dots, n-1, r_0 = r_n = 1, \overline{r} = \max\{r_\ell\} \Big\}. \end{aligned} \tag{11}$$

Note that the set (11) contains not only PSD matrices but also non-PSD matrices. Indeed, one may impose additional structure, such as [33, eq. (3)], on the factors $\{\boldsymbol{X}_\ell^{i_\ell, j_\ell}\}$ to ensure $\boldsymbol{\rho}$ is PSD. However, the condition in [33, eq. (3)] is only sufficient rather than necessary for ensuring $\boldsymbol{\rho}$ is PSD. Moreover, adding the PSD constraint does not significantly reduce the number of degrees of freedom of elements in the set $\mathbb{X}_{\overline{r}}$. Therefore, in the following, we will simply focus on the set of generic MPOs (11) without a PSD constraint.

Efficiency of MPO representation Due to the curse of dimensionality, the number of elements in the density matrix $\boldsymbol{\rho}$ grows exponentially in the number of qudits n. In contrast, the MPO form (7) can represent $\boldsymbol{\rho}$ using only $O(nd^2\overline{r}^2)$ elements, where $\overline{r} = \max\{r_1, \dots, r_{n-1}\}$. This makes the MPO form remarkably effective in combatting the curse of dimensionality as its number of parameters scales only linearly in terms of n. The concise representation provided by MPO is remarkably useful in QST since it may allow us to reconstruct a quantum state with both experimental and computational resources that are only *polynomial* rather than *exponential* in the number of qudits [75–78]. Beyond applications in quantum information processing, the equivalent form of TT decomposition mentioned above has also been widely used for image compression [58,79], analyzing theoretical properties of deep networks [80], network compression (or tensor networks) [81–86], recommendation systems [87], probabilistic model estimation [88], and learning of Hidden Markov Models [89] to mention a few usages.[7]

Linear combination of MPOs In linear algebra, the (matrix) rank of the sum of two matrices is less than or equal to the sum of the (matrix) ranks of these matrices. This also holds for MPO ranks. In particular, for any two MPOs $\widetilde{\boldsymbol{\rho}}, \widehat{\boldsymbol{\rho}} \in \mathbb{C}^{d^n \times d^n}$ of the form (7) with factors $\{\widetilde{\boldsymbol{X}}^{i_\ell, j_\ell} \in \mathbb{C}^{\widetilde{r}_{\ell-1} \times \widetilde{r}_\ell}\}$ and $\{\widehat{\boldsymbol{X}}^{i_\ell, j_\ell} \in \mathbb{C}^{\widehat{r}_{\ell-1} \times \widehat{r}_\ell}\}$, respectively, the elements of their summation $\boldsymbol{\rho} = \widetilde{\boldsymbol{\rho}} + \widehat{\boldsymbol{\rho}}$ can be expressed by

$$\boldsymbol{\rho}(i_1 \cdots i_n, j_1 \cdots j_n) = \begin{bmatrix} \widetilde{\boldsymbol{X}}_1^{i_1, j_1} & \widehat{\boldsymbol{X}}_1^{i_1, j_1} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{X}}_2^{i_2, j_2} & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{X}}_2^{i_2, j_2} \end{bmatrix} \cdots \begin{bmatrix} \widetilde{\boldsymbol{X}}_{n-1}^{i_{n-1}, j_{n-1}} & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{X}}_{n-1}^{i_{n-1}, j_{n-1}} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{X}}_n^{i_n, j_n} \\ \widehat{\boldsymbol{X}}_n^{i_n, j_n} \end{bmatrix}, \tag{12}$$

implying that the MPO ranks r_ℓ of $\boldsymbol{\rho}$ satisfy $r_\ell \le \widehat{r}_\ell + \widetilde{r}_\ell$ for all $\ell = 1, \dots, n-1$.

[6] The right-canonical form refers to the case where $R(\boldsymbol{X}_\ell)$ are unitary matrices for all $\ell = 2, \dots, n$.

[7] See [90] for a python library for TT decomposition.

3 Stable Embeddings of Matrix Product Operators

3.1 Background

Measurements must satisfy certain properties to enable recovery of quantum states. One desirable property known as a *stable embedding* has been widely studied and popularized in the compressive sensing literature [5,38–41]. In this section, we will study the embedding of MPOs from various measurement types including quantum measurements. Towards that goal, we will first consider population measurements, and in the next section, we will study stable recovery with empirical measurements.

As described in Section 2.2, the population measurements from one POVM are linear measurements that can be described through a linear map $\mathcal{A}:\mathbb{C}^{d^n\times d^n}\to\mathbb{R}^K$ of the form

$$\mathcal{A}(\boldsymbol{\rho})=\begin{bmatrix}\langle\boldsymbol{A}_1,\boldsymbol{\rho}\rangle\\ \vdots\\ \langle\boldsymbol{A}_K,\boldsymbol{\rho}\rangle\end{bmatrix}. \tag{13}$$

According to the discussion in Section 2.2, the choice of $\{\boldsymbol{A}_k\}$ can vary. Our goal is to study the properties of the associated measurement operators.

Our study of stable embeddings of MPOs from population measurements concerns the quantity $\|\mathcal{A}(\boldsymbol{\rho})\|_2^2$. As described in Section 3.2, a favorable situation is when $\mathcal{A}$ satisfies the restricted isometry property (RIP), where $\|\mathcal{A}(\boldsymbol{\rho})\|_2^2$ is guaranteed to be proportional to $\|\boldsymbol{\rho}\|_F^2$ for any MPO $\boldsymbol{\rho}$. In some cases, only a lower bound on this proportionality can be established. In particular, in Section 3.3, we establish a guarantee of the form

$$\|\mathcal{A}(\boldsymbol{\rho})\|_2^2\geq C_{d,n,K}\|\boldsymbol{\rho}\|_F^2, \tag{14}$$

where $C_{d,n,K}$ is a positive constant depending on d,n,K, and the guarantee holds uniformly for all MPOs up to some maximum rank. When this holds, then for any two MPOs $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$, noting that $\boldsymbol{\rho}_1-\boldsymbol{\rho}_2$ is also an MPO according to (12), we have

$$\|\mathcal{A}(\boldsymbol{\rho}_1)-\mathcal{A}(\boldsymbol{\rho}_2)\|_2^2\geq C_{d,n,K}\left\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\right\|_F^2,$$

which ensures distinct measurements (i.e., $\mathcal{A}(\boldsymbol{\rho}_1)\neq\mathcal{A}(\boldsymbol{\rho}_2)$) as long as $\boldsymbol{\rho}_1\neq\boldsymbol{\rho}_2$.

In compressive sensing of sparse signals and low-rank matrices [5,38–41], uniform stable embeddings of all possible signals of interest can often be achieved by choosing the measurement operators randomly from a certain distribution. Thus, random matrices and projections have played a central role in the analysis of the associated inverse problems [42]. In this section, we will study the embeddings of MPOs from linear measurements where the measurement matrices $\{\boldsymbol{A}_k\}$ are generated from certain random distributions. Specifically, we will first study perhaps the most generic random distribution where all the elements of $\boldsymbol{A}_k$ are independently generated from a Gaussian distribution. We will then study rank-one random POVM measurements of the form $\boldsymbol{A}_k=\boldsymbol{a}_k\boldsymbol{a}_k^H$ with each $\boldsymbol{a}_k$ randomly generated from a multivariate normal distribution. Finally, we will study the physically realizable (though inefficient) measurements acquired using multiple Haar random rank-one POVMs.

Normalized set of MPOs Since $\mathcal{A}(\cdot)$ is a linear map, without loss of generality, we will focus on MPOs $\boldsymbol{\rho}\in\mathbb{X}_{\overline{r}}$ with unit Frobenius norm. By the left-canonical form in (10), we have

$$\begin{aligned}
\|\boldsymbol{\rho}\|_F^2&=\sum_{i_1,j_1}\cdots\sum_{i_n,j_n}\left(\boldsymbol{X}_n^{i_n,j_n}\right)^H\cdots\left(\boldsymbol{X}_1^{i_1,j_1}\right)^H\boldsymbol{X}_1^{i_1,j_1}\cdots\boldsymbol{X}_n^{i_n,j_n}\\
&=\sum_{i_2,j_2}\cdots\sum_{i_n,j_n}\left(\boldsymbol{X}_n^{i_n,j_n}\right)^H\cdots\left(\boldsymbol{X}_2^{i_2,j_2}\right)^H\underbrace{\left(\sum_{i_1,j_1}\left(\boldsymbol{X}_1^{i_1,j_1}\right)^H\boldsymbol{X}_1^{i_1,j_1}\right)}_{\mathbf{I}_{r_1}}\boldsymbol{X}_2^{i_2,j_2}\cdots\boldsymbol{X}_n^{i_n,j_n}\\
&=\sum_{i_2,j_2}\cdots\sum_{i_n,j_n}\left(\boldsymbol{X}_n^{i_n,j_n}\right)^H\cdots\left(\boldsymbol{X}_2^{i_2,j_2}\right)^H\boldsymbol{X}_2^{i_2,j_2}\cdots\boldsymbol{X}_n^{i_n,j_n}=\cdots\\
&=\sum_{i_n,j_n}{\boldsymbol{X}_n^{i_n,j_n}}^H\boldsymbol{X}_n^{i_n,j_n},
\end{aligned}$$

which together with $\|\boldsymbol{\rho}\|_F^2 = 1$ also leads to $\sum_{i_n,j_n}\left(\boldsymbol{X}_n^{i_n,j_n}\right)^H \boldsymbol{X}_n^{i_n,j_n} = 1$. Thus, the set of all the MPOs $\boldsymbol{\rho} \in \mathbb{X}_{\overline{r}}$ with unit norm, denoted by $\overline{\mathbb{X}}_{\overline{r}}$, can also be expressed by

$$\begin{aligned}\overline{\mathbb{X}}_{\overline{r}} = \Big\{&\boldsymbol{\rho} \in \mathbb{C}^{d^n \times d^n} : \boldsymbol{\rho} = \boldsymbol{\rho}^H, \boldsymbol{\rho}(i_1 \cdots i_n, j_1 \cdots j_n) = \boldsymbol{X}_1^{i_1,j_1}\boldsymbol{X}_2^{i_2,j_2}\cdots\boldsymbol{X}_n^{i_n,j_n}, \boldsymbol{X}_\ell^{i_\ell,j_\ell} \in \mathbb{C}^{r_{\ell-1}\times r_\ell}, \\ &\sum_{i_\ell,j_\ell}\left(\boldsymbol{X}_\ell^{i_\ell,j_\ell}\right)^H \boldsymbol{X}_\ell^{i_\ell,j_\ell} = \mathbf{I}_{r_\ell}, \ell = 1,\dots,n, r_0 = r_n = 1, \overline{r} = \max\{r_\ell\}\Big\}.\end{aligned} \tag{15}$$

3.2 Restricted isometry property with generic Gaussian measurements

To provide a baseline for the sample complexity of population measurements, we begin by studying perhaps the most generic type of random measurements, where each entry of $\boldsymbol{A}_k$ is i.i.d. standard complex Gaussian random variable $X = \mathscr{R}(X) + i\mathscr{I}(X)$ with $\mathscr{R}(X)$ and $\mathscr{I}(X)$ being independent and following $\mathcal{N}(0, \frac{1}{2})$, the Gaussian distribution with mean 0 and variance $\frac{1}{2}$. Such measurements do not form a POVM and thus cannot be physically implemented in quantum measurement systems. However, as Gaussian measurements provide the "gold standard" for random linear measurement operators in many compressive sensing and low-rank matrix recovery problems, their sample complexity for stable embeddings of MPOs provides useful insight.

Gaussian measurements can be shown to satisfy a strong type of stable embedding guarantee known as the restricted isometry property (RIP).

Definition 2 (Restricted isometry property (RIP))**.** *A linear operator $\mathcal{A} : \mathbb{C}^{d^n \times d^n} \to \mathbb{C}^K$ is said to satisfy the $\delta_{\overline{r}}$-restricted isometry property ($\delta_{\overline{r}}$-RIP) if*

$$(1 - \delta_{\overline{r}})\|\boldsymbol{\rho}\|_F^2 \leq \frac{1}{K}\|\mathcal{A}(\boldsymbol{\rho})\|_2^2 \leq (1 + \delta_{\overline{r}})\|\boldsymbol{\rho}\|_F^2 \tag{16}$$

holds for any density operator $\boldsymbol{\rho} \in \mathbb{C}^{d^n \times d^n}$ which has the MPO format with MPO ranks $\boldsymbol{r} = (r_1, \dots, r_{n-1}), r_i \leq \overline{r}$.

The following result establishes the RIP for Gaussian measurements.

Theorem 2. *Suppose that each entry of $\boldsymbol{A}_k$ in the linear map $\mathcal{A} : \mathbb{C}^{d^n \times d^n} \to \mathbb{C}^K$ defined in* (13) *is an i.i.d. standard complex Gaussian random variable. Then, with probability at least $1 - \overline{\epsilon}$, $\mathcal{A}$ satisfies the $\delta_{\overline{r}}$-RIP as in* (16) *for MPOs given that*

$$K \geq C \cdot \frac{1}{\delta_{\overline{r}}^2} \cdot \max\left\{nd^2\overline{r}^2(\log n\overline{r}), \log(1/\overline{\epsilon})\right\}, \tag{17}$$

where C is a universal constant.

In Appendix A, we extended this result to generic subgaussian measurements. We note that a similar result for TT-format tensors in the real domain was given in [71], and we share similar techniques for proving the RIP by using tools involving the ϵ-net and covering arguments [91,92] and deviation bounds for the supremum of a chaos process [93,94]. While MPOs are equivalent in form to TT-format tensors as discussed in Section 2.3, we provide the proof in Appendix A for the sake of completeness and because here we consider the complex domain. Also, the sampling complexity in [71] is $K \gtrsim \frac{1}{\delta_{\overline{r}}^2} \cdot \max\left\{((n-1)\overline{r}^3 + nd^2\overline{r})(\log n\overline{r}), \log(1/\overline{\epsilon})\right\}$, which is slightly different from (17). Considering a qubit system with $d = 2$, the main order $n\overline{r}^2$ in (17) is slightly better than the order $(n-1)\overline{r}^3$ from [71] when the bond dimension $\overline{r}$ is large.

Although the Gaussian measurements are not POVMs and cannot be directly used for quantum measurements, Theorem 2 indicates that it is possible to estimate an MPO state with $\widetilde{\Omega}(nd^2\overline{r}^2)$ linear measurements. In comparison, for a state with low (matrix) rank structure, say rank r, $\widetilde{\Omega}(d^n r)$ measurements are needed even with Gaussian measurements [6].

3.3 Stable embeddings with rank-one POVM measurements

We now study the population measurements arising from structured rank-one measurement ensembles with PSD matrices $\boldsymbol{A}_k = \boldsymbol{\psi}_k\boldsymbol{\psi}_k^H$ as introduced in Section 2.2. We first consider the case where we omit the constraint (1) that the matrices $\boldsymbol{A}_k$ sum to the identity matrix. Rather, we simply generate the $\boldsymbol{\psi}_k = \boldsymbol{a}_k$ independently and randomly from a certain distribution, specifically, $\boldsymbol{a}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d^n})$. The independence among $\{\boldsymbol{a}_k\}$ will simplify the analysis and help derive a tight bound for stable embedding. We call such measurements *rank-one independent POVM measurements*. We then consider the practical case ($\boldsymbol{\psi}_k = \boldsymbol{\phi}_k$) where $\{\boldsymbol{\phi}_k\}$ are generated from a Haar-distributed random unitary matrix, which results in *Haar random rank-one POVM measurements*.

Rank-one Gaussian measurements It is known that rank-one measurements do not obey the RIP condition for low-rank matrices [19,46]. Since we expect this to also be true for MPOs, we instead aim to establish a lower bound on the isometry of the form (14). Towards that goal, we will use Mendelson's small ball method [42,47,48] for establishing a lower bound on a nonnegative empirical process.

Lemma 1. *([42,47,48]) Fix a set $E \subset \mathbb{C}^D$. Let $\boldsymbol{b}$ be a random vector on $\mathbb{C}^D$ and let $\boldsymbol{b}_1, \dots, \boldsymbol{b}_K$ be independent copies of $\boldsymbol{b}$. Introduce the marginal tail function*

$$H_\xi(E; \boldsymbol{b}) = \inf_{\boldsymbol{u} \in E} \mathbb{P}\{|\langle \boldsymbol{b}, \boldsymbol{u} \rangle| \geq \xi\}, \textit{ for } \xi > 0. \tag{18}$$

Let $\epsilon_k, k = 1, \dots, K$, be independent Rademacher random variables, independent from everything else. Define the mean empirical width of the set E as

$$W_K(E; \boldsymbol{b}) = \mathbb{E} \sup_{\boldsymbol{u} \in E} \langle \boldsymbol{h}, \boldsymbol{u} \rangle, \textit{ where } \boldsymbol{h} = \frac{1}{\sqrt{K}} \sum_{k=1}^K \epsilon_k \boldsymbol{b}_k. \tag{19}$$

Then, for any $\xi > 0$ and $t > 0$, with probability at least $1 - e^{-\frac{t^2}{2}}$ we have

$$\inf_{\boldsymbol{u} \in E} \left(\sum_{k=1}^K |\langle \boldsymbol{b}_k, \boldsymbol{u} \rangle|^2 \right)^{\frac{1}{2}} \geq \xi \sqrt{K} H_\xi(E; \boldsymbol{b}) - 2W(E; \boldsymbol{b}) - t\xi. \tag{20}$$

This result delivers an effective lower bound for a nonnegative empirical process defined in the left-hand side of (20). This result is also utilized for studying stable embeddings for low-rank matrices [19,95]. Noting the similar forms between (20) and (14), we apply Lemma 1 for our case where the set E becomes $\overline{\mathbb{X}}_{\overline{r}}$ and $\boldsymbol{b}$ becomes a random measurement matrix of form $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}^H$ with $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d^n})$. We then need to analyze the following marginal tail function and mean empirical width

$$\begin{aligned} H_\xi(\overline{\mathbb{X}}_{\overline{r}}; \boldsymbol{A}) &= \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \mathbb{P}\{|\langle \boldsymbol{A}, \boldsymbol{\rho} \rangle| \geq \xi\}, \\ W_K(\overline{\mathbb{X}}_{\overline{r}}; \boldsymbol{A}) &= \frac{1}{\sqrt{K}} \mathbb{E} \sup_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \sum_{k=1}^K \langle \epsilon_k \boldsymbol{A}_k, \boldsymbol{\rho} \rangle. \end{aligned}$$

As in [19,42], we can use the Payley-Zygmund inequality to obtain a lower bound for the marginal tail function $H_\xi(\overline{\mathbb{X}}_{\overline{r}}; \boldsymbol{A})$. In terms of the mean empirical width $W_K(\overline{\mathbb{X}}_{\overline{r}}; \boldsymbol{A})$, the work [19,42] uses an inequality that directly upper bounds the supremum of $\langle \boldsymbol{A}, \boldsymbol{\rho} \rangle$ over rank-r matrices $\boldsymbol{\rho}$ by $2\sqrt{r}\|\boldsymbol{A}\|$. Unfortunately, it is difficult to extend this approach to our case. Instead, we use an ϵ-netargument to provide a uniform upper bound for $\langle \boldsymbol{A}, \boldsymbol{\rho} \rangle$. With the detailed analysis in Appendix B, we establish the following result.

Theorem 3. *Let $\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_K\}$ be selected independently and randomly from the multivariate standard normal distribution $\boldsymbol{I}_{d^n}$. Given*

$$K \gtrsim nd^2\overline{r}^2 \log n, \tag{21}$$

then the induced linear map $\mathcal{A}$ with measurement operators $\{\boldsymbol{A}_k = \boldsymbol{a}_k\boldsymbol{a}_k^H\}$ satisfies

$$\|\mathcal{A}(\boldsymbol{\rho})\|_2 \gtrsim \sqrt{K}, \ \forall \boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}} \tag{22}$$

with probability at least $1 - e^{-\alpha_1 K}$, where α_1 is a positive constant.

Under the same setup, one requires $K \gtrsim d^n r$ measurement operators for the induced linear map $\mathcal{A}$ to obey the stable embedding property for rank-r matrices [19]. Fortunately, due to the extremely low-dimensional structure of the MPO format, the number of measurement operators only needs to scale linearly in terms of the number of qudits n (if we ignore the logarithmic term).

Haar random rank-one POVM measurements We now study practical measurements consisting of an ensemble of Haar random rank-one POVMs as described in Section 2.2. Let $\begin{bmatrix}\boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n}\end{bmatrix}, i = 1, \dots, Q$ be Q randomly generated Haar-distributed unitary matrices. According to Section 2.2, each unitary matrix induces a linear operator $\mathcal{A}_i : \mathbb{C}^{d^n \times d^n} \to \mathbb{R}^K$ that generates population measurements for a quantum state $\boldsymbol{\rho}$ as

$$\mathcal{A}_i(\boldsymbol{\rho}) = \begin{bmatrix} \langle \boldsymbol{A}_{i,1}, \boldsymbol{\rho} \rangle \\ \vdots \\ \langle \boldsymbol{A}_{i,K}, \boldsymbol{\rho} \rangle \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{\phi}_{i,1}\boldsymbol{\phi}_{i,1}^H, \boldsymbol{\rho} \rangle \\ \vdots \\ \langle \boldsymbol{\phi}_{i,K}\boldsymbol{\phi}_{i,K}^H, \boldsymbol{\rho} \rangle \end{bmatrix}, \tag{23}$$

where in practice we will use $K = d^n$, but for generality we can choose any $K \leq d^n$. We note that for each i, even though $\begin{bmatrix}\boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n}\end{bmatrix}$ is unitary and $\sum_{k=1}^{d^n} \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H = \mathbf{I}$, $\mathcal{A}_i$ is not an identity mapping in $\mathbb{C}^{d^n \times d^n}$ even with $K = d^n$; this is because $\mathcal{A}_i$ collects at most d^n measurements of an object $\boldsymbol{\rho}$ that contains d^{2n} entries. We now stack all the population measurements together as

$$\mathcal{A}^Q(\boldsymbol{\rho}) = \begin{bmatrix} \mathcal{A}_1(\boldsymbol{\rho}) \\ \vdots \\ \mathcal{A}_Q(\boldsymbol{\rho}) \end{bmatrix}, \tag{24}$$

where $\mathcal{A}^Q : \mathbb{C}^{d^n \times d^n} \to \mathbb{R}^{KQ}$ denotes the linear operator corresponding to the Q POVMs.

For any i, since $\boldsymbol{\phi}_{i,k}$ and $\boldsymbol{\phi}_{i,k'}$ may not be independent for any $k \neq k'$, we cannot directly apply Lemma 1 to study stable embeddings via $\mathcal{A}^Q$. To address this issue, we modify Mendelson's small ball method as follows.

Lemma 2. *Consider a fixed set $E \subset \mathbb{C}^D$. Let $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_K\}$ represent a collection of random columns in $\mathbb{C}^D$, which may not be mutually independent. Additionally, let $\{\boldsymbol{b}_{i,1}, \dots, \boldsymbol{b}_{i,K}\}_{i=1}^Q$ denote a set of independent copies of $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_K\}$. Introduce the marginal tail function*

$$H_\xi(E; \boldsymbol{b}) = \inf_{\boldsymbol{u} \in E} \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\{|\langle \boldsymbol{b}_k, \boldsymbol{u} \rangle| \geq \xi\}, \ for\ \xi > 0. \tag{25}$$

Let $\epsilon_i, i = 1, \dots, Q$ be independent Rademacher random variables, independent from everything else, and define the mean empirical width of the set:

$$W_{QK}(E; \boldsymbol{b}) = \mathbb{E} \sup_{\boldsymbol{u} \in E} \langle \boldsymbol{h}, \boldsymbol{u} \rangle, \ where\ \boldsymbol{h} = \frac{1}{\sqrt{QK}} \sum_{i=1}^{Q} \sum_{k=1}^{K} \epsilon_i \boldsymbol{b}_{i,k}. \tag{26}$$

Then, for any $\xi > 0$ and $t > 0$

$$\inf_{\boldsymbol{u} \in E} \left(\sum_{i=1}^{Q} \sum_{k=1}^{K} |\langle \boldsymbol{b}_{i,k}, \boldsymbol{u} \rangle|^2 \right)^{\frac{1}{2}} \geq \xi \sqrt{QK} H_\xi(E; \boldsymbol{b}) - 2W_{QK}(E; \boldsymbol{b}) - t\xi\sqrt{K}, \tag{27}$$

with probability at least $1 - e^{-\frac{t^2}{2}}$.

The proof has been provided in Appendix C. Note that when $K = 1$, Lemma 2 reduces to Lemma 1 (by setting $Q = K$ in Lemma 2). In other words, Q plays the same role as K in Lemma 1. To effectively apply the modified method, we need to generalize the linear map in (13). With Lemma 2, we now establish the stable embedding of (24) in the following theorem.

Theorem 4 (Stable embedding of multiple Haar random rank-one POVMs). *Let $\mathcal{A}^Q : \mathbb{C}^{d^n \times d^n} \to \mathbb{R}^{KQ}$ be the linear mapping defined in* (24) *that is induced by Q random unitary matrices. For any $K \geq 1$, assuming*

$$Q \gtrsim nd^2\overline{r}^2 \log n, \ \overline{r} = \max_{i=1,\dots n-1} r_i, \tag{28}$$

then with probability at least $1 - e^{-\alpha_2 Q}$ (where α_2 is a positive constant.), $\mathcal{A}^Q$ obeys

$$\|\mathcal{A}^Q(\boldsymbol{\rho})\|_2 \gtrsim \frac{\sqrt{QK}}{d^n} \tag{29}$$

for any $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$.

The proof is given in Appendix D. First note that in Theorem 4, the requirement on Q in (28) and the failure probability $e^{-\alpha_3 Q}$ are similar to those in Theorem 3 on K. This is because, as we explained before, Q in Lemma 2 plays the same role as K in Lemma 1, and likewise Q in Theorem 4 is equivalent to K in (21). Thus, Theorem 4 holds for any $K \geq 1$. On the other hand, without exploiting the randomness between different columns within a random unitary matrix, the number of POVMs Q is required to be relatively large as stated in (28). Considering that the local correlations between the columns in the unitary matrix are very weak because the orthogonality is a global property [96], we conjecture that the requirement on Q can be significantly reduced, even to $Q = 1$. Indeed, according to [97, Theorem 3], when $n \to \infty$, in an "in probability" sense, all elements (scaled by $\sqrt{d^n}$) of $o(\frac{d^n}{n \log d})$ columns in a Haar-distributed random unitary matrix can be approximated by entries generated independently from a standard normal distribution. As $o(\frac{d^n}{n \log d})$ independent columns from a multivariate normal distribution are sufficient for Theorem 3, this suggests that it is highly possible to ensure stable embedding (29) with a single POVM $Q = 1$. While we leave a formal analysis as future work, we conduct a numerical experiment to support this conjecture. Set $d = 2, Q = 1, K = d^n, r_1 = \cdots = r_{n-1} = 2$. Then for each n, we randomly generate a unitary matrix (i.e., $Q = 1$), randomly sample many MPOs $\boldsymbol{\rho}$ with $\|\boldsymbol{\rho}\|_F = 1$, and compute the minimum of $\|\mathcal{A}^Q(\boldsymbol{\rho})\|_2$ among all the generated MPOs. In Fig. 2, we compare the minimum of $\|\mathcal{A}^Q(\boldsymbol{\rho})\|_2$ (averaged over 50 Monte Carlo trails) with $\frac{1}{\sqrt{d^n}}$. We observe that $\|\mathcal{A}^Q(\boldsymbol{\rho})\|_2$ is of the same order as $\frac{1}{\sqrt{d^n}}$. Furthermore, as the number of qudits increases, $\|\mathcal{A}^Q(\boldsymbol{\rho})\|_2$ approaches $\frac{1}{\sqrt{d^n}}$. This is consistent with (29), where the right hand side becomes $\frac{1}{\sqrt{d^n}}$ when $K = d^n, Q = 1$.

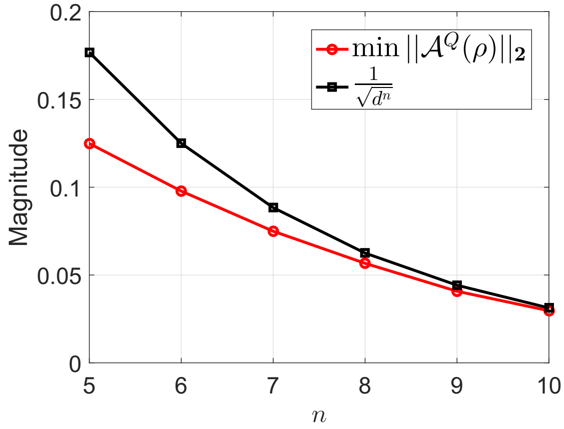


Figure 2: Numerical computation of $\min_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \|\mathcal{A}^Q(\boldsymbol{\rho})\|_2$ with $Q = 1$ and $K = d^n$.

4 Stable recovery with empirical measurements

The results of Section 3.3 ensure a distinct set of population measurements $\mathcal{A}^Q(\boldsymbol{\rho})$ for any ground-truth MPO $\boldsymbol{\rho}^\star$ under multiple Haar random rank-one POVM measurements. Based on these results, in this section, we study the stable recovery of $\boldsymbol{\rho}$ from empirical measurements obtained by multiple Haar random rank-one POVMs. With Q randomly generated Haar-distributed unitary matrices $\begin{bmatrix} \boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n} \end{bmatrix}, i = 1, \dots, Q$, according to (24), we can generate Qd^n population measurements through the linear measurement operator $\mathcal{A}^Q : \mathbb{C}^{d^n \times d^n} \to \mathbb{R}^{Qd^n}$ (set $K = d^n$ in (24)) as

$$\boldsymbol{p}^Q = \mathcal{A}^Q(\boldsymbol{\rho}^\star) = \begin{bmatrix} \boldsymbol{p}_1 \\ \vdots \\ \boldsymbol{p}_Q \end{bmatrix} = \begin{bmatrix} \mathcal{A}_1(\boldsymbol{\rho}^\star) \\ \vdots \\ \mathcal{A}_Q(\boldsymbol{\rho}^\star) \end{bmatrix}, \tag{30}$$

where $\mathcal{A}_i$ is as defined in (23) with $K = d^n$ and with $\boldsymbol{A}_{i,k} = \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H$. Denote by $p_{i,k}$ the k-th element in $\boldsymbol{p}_i$.

For each POVM, suppose we repeat the measurement process M times and take the average of the outcomes to generate empirical probabilities

$$\widehat{p}_{i,k} = \frac{f_{i,k}}{M},\ i = 1,\dots,Q,\ k = 1,\dots,d^n, \tag{31}$$

where $f_{i,k}$ denotes the number of times the k-th output is observed when using the i-th POVM M times. Denote by $\widehat{\boldsymbol{p}}_i = \begin{bmatrix}\widehat{p}_{i,1} & \cdots & \widehat{p}_{i,d^n}\end{bmatrix}^\top$ the empirical measurements obtained by the i-th POVM and stack all the total empirical measurements together as $\widehat{\boldsymbol{p}}^Q = \begin{bmatrix}\widehat{\boldsymbol{p}}_1^T & \cdots & \widehat{\boldsymbol{p}}_Q^T\end{bmatrix}^\top$, which are unbiased estimators of the population measurements $\boldsymbol{p}^Q$. We denote by $\boldsymbol{\eta}$ the measurement error as

$$\boldsymbol{\eta} = \widehat{\boldsymbol{p}}^Q - \boldsymbol{p}^Q = \widehat{\boldsymbol{p}}^Q - \mathcal{A}^Q(\boldsymbol{\rho}^\star) = \left[\boldsymbol{\eta}_1^T, \cdots, \boldsymbol{\eta}_Q^T\right]^T, \tag{32}$$

where $\eta_{i,k}$ is the k-th element in $\boldsymbol{\eta}_i$.

With empirical measurements $\widehat{\boldsymbol{p}}^Q$, for simplicity, we consider minimizing the following constrained least squares objective:

$$\widehat{\boldsymbol{\rho}} = \operatorname*{arg\,min}_{\boldsymbol{\rho}\in\mathbb{X}_{\overline{r}}} \|\mathcal{A}^Q(\boldsymbol{\rho}) - \widehat{\boldsymbol{p}}^Q\|_2^2, \tag{33}$$

where $\mathcal{A}^Q$ is the induced linear map as defined in (30). Supposing one can find a global solution of (33), our goal is to study how the recovery error $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F$ scales with the size of the MPO (particularly with respect to the number of qudits n) and the total number of measurements QM. To enable a stable estimate of the state by measuring it only a polynomial number of times in terms of n, we desire the recovery error to grow only polynomially rather than exponentially in n.

4.1 Challenge: Abundant but extremely noisy measurements

Before presenting the main result, we first discuss the challenge and hope of obtaining a recovery error that only grows polynomially in terms of the number of qudits. Recall that $(f_{i,1},\dots,f_{i,d^n})$ in (31) follows a multinomial distribution $\mathrm{Multinomial}(M, \boldsymbol{p}_i)$ with parameters M and $\boldsymbol{p}_i$. Thus, $\widehat{\boldsymbol{p}}_i - \boldsymbol{p}_i$ has mean zero and covariance matrix $\boldsymbol{\Sigma}_i$, where $\boldsymbol{\Sigma}_i$ has elements given by $\boldsymbol{\Sigma}_i[l,j] = \begin{cases} \frac{p_{i,l}(1-p_{i,l})}{M}, & l = j \\ -\frac{p_{i,l}p_{i,j}}{M}, & l \neq j \end{cases}$. With this observation, we have

$$\mathbb{E}\,\|\boldsymbol{\eta}\|_2^2 = \mathbb{E}\left[\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}^2\right] = \sum_{i=1}^{Q}\sum_{k=1}^{d^n}\frac{p_{i,k}(1-p_{i,k})}{M} \leq \frac{Q}{M}. \tag{34}$$

Note that $\frac{p_{i,k}(1-p_{i,k})}{M}$ could be as small as 0 which can be achieved, although rarely, when $p_{i,k} \in \{0,1\}$ (i.e., when $\{p_{i,k}, k = 1,\dots,d^n\}$ has a spiky distribution). However, the above bound on the order of $\frac{Q}{M}$ is tight when the distribution of $\{p_{i,k}, k = 1,\dots,d^n\}$ is not spiky (e.g., when each $p_{i,k}$ is on the order of $\frac{1}{d^n}$). To see this, denote the eigenvalue decomposition of $\boldsymbol{\rho}^\star$ as $\boldsymbol{\rho}^\star = \sum_{i=1}^{d^n}\lambda_i\boldsymbol{u}_i\boldsymbol{u}_i^H$, where $\sum_{i=1}^{d^n}\lambda_i = 1$. Now for any i and k, we can compute $\mathbb{E}[p_{i,k}^2]$ as

$$\begin{aligned}
\mathbb{E}[|\langle\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}^\star\rangle|^2] &= \sum_{j=1}^{d^n}\sum_{l=1}^{d^n}\lambda_j\lambda_l\,\mathbb{E}[|\boldsymbol{\phi}_{i,k}^H\boldsymbol{u}_j|^2|\boldsymbol{\phi}_{i,k}^H\boldsymbol{u}_l|^2] \\
&= \sum_{l\neq j}\lambda_j\lambda_l(\mathbb{E}[|\boldsymbol{\phi}_{i,k}[1]|^2])^2 + \sum_{l}\lambda_l^2\,\mathbb{E}[|\boldsymbol{\phi}_{i,k}[1]|^4] \\
&= \sum_{l\neq j}\frac{\lambda_j\lambda_l}{d^{2n}} + 2\sum_{l}\frac{\lambda_l^2}{d^n(d^n+1)} \\
&= \sum_{j=1}^{d^n}\sum_{l=1}^{d^n}\frac{\lambda_j\lambda_l}{d^{2n}} + \sum_{l=1}^{d^n}\frac{d^n-1}{d^{2n}(d^n+1)}\lambda_l^2 \\
&= \frac{1}{d^{2n}} + \frac{d^n-1}{d^{2n}(d^n+1)}\|\boldsymbol{\rho}^\star\|_F^2,
\end{aligned} \tag{35}$$

where $\boldsymbol{\phi}_{i,k}[1]$ is the first element of $\boldsymbol{\phi}_{i,k}$, the second line utilizes the rotation invariance of the unitary matrix in Lemma 8, and the third line uses Lemma 9.

Noting that $\|\boldsymbol{\rho}^\star\|_F^2 \leq (\sum_{i=1}^{d^n} \lambda_i)^2 = 1$, we further have

$$\frac{1}{d^{2n}} \leq \mathbb{E}[p_{i,k}^2] \leq \frac{2}{d^{2n}},\ \forall 1 \leq i \leq Q \ \text{ and } \ \forall 1 \leq k \leq d^n. \tag{36}$$

In other words, if $\begin{bmatrix}\boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n}\end{bmatrix}$ is a randomly generated unitary matrix, then each $p_{i,k}$ has the same second moment of order $1/d^{2n}$. This suggests that the distribution of $\{p_{i,k}, k = 1, \dots, d^n\}$ is more uniform than spiky.

In addition, (36) also gives the energy of the clean measurements or population measurements as

$$\frac{Q}{d^n} \leq \mathbb{E}\left[\sum_{i=1}^{Q}\sum_{k=1}^{d^n} p_{i,k}^2\right] = \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \mathbb{E}\langle \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}^\star\rangle^2 \leq \frac{2Q}{d^n}. \tag{37}$$

To summarize, the above discussion gives the following comparison between the energy of the clean measurements and the noise in the measurements:

$$\text{Clean measurements: } \mathbb{E}\,\|\boldsymbol{p}^Q\|_2^2 = O\left(\frac{Q}{d^n}\right),$$
$$\text{Statistical error: } \mathbb{E}\,\|\boldsymbol{\eta}\|_2^2 = O\left(\frac{Q}{M}\right),$$

which indicates that the statistical error or measurement noise is exponentially larger than the clean measurements. This seems to suggest that M has to be on the order of d^n to obtain measurements with suitable signal-to-noise ratio for stable recovery.

Fortunately, though each measurement could be extremely noisy, we have an exponentially large number of such measurements $\{\widehat{p}_{i,1}, \dots, \widehat{p}_{i,d^n}\}_i$, from Q POVMs. This setting is slightly different from some common inverse problems [42,98,99], where the number of measurements matches the number of degrees of freedom behind the underlying signal but the measurements are not overwhelmed by noise. In addition, conditioned on the selected POVM, the measurement noise $\boldsymbol{\eta}$ is random and behaves close to a multivariate Gaussian distribution [100–102]. By exploiting these observations together with the stable embeddings established in the last section, we anticipate stable recovery even when M is only polynomially large in n.

4.2 Stable recovery with empirical measurements

We now provide a formal analysis of the recovery error $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F$, where $\widehat{\boldsymbol{\rho}}$ is a global solution of (33). Using (33) and the fact that $\boldsymbol{\rho}^\star \in \mathbb{X}_{\overline{r}}$, we have

$$\begin{aligned} 0 \ &\leq \ \|\mathcal{A}^Q(\boldsymbol{\rho}^\star) - \widehat{\boldsymbol{p}}^Q\|_2^2 - \|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}}) - \widehat{\boldsymbol{p}}^Q\|_2^2 \\ &= \ \|\mathcal{A}^Q(\boldsymbol{\rho}^\star) - \mathcal{A}^Q(\boldsymbol{\rho}^\star) - \boldsymbol{\eta}\|_2^2 - \|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}}) - \mathcal{A}^Q(\boldsymbol{\rho}^\star) - \boldsymbol{\eta}\|_2^2 \\ &= \ 2\langle \mathcal{A}^Q(\boldsymbol{\rho}^\star) + \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle + \|\mathcal{A}^Q(\boldsymbol{\rho}^\star)\|_2^2 - \|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}})\|_2^2 \\ &= \ 2\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle - \|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\|_2^2, \end{aligned} \tag{38}$$

which further implies that

$$\|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\|_2^2 \leq 2\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle. \tag{39}$$

The left-hand side of the above equation can be further lower bounded by order of $\frac{Q}{d^n}\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F^2$ according to Theorem 4. The challenging part is to deal with the right-hand side of (39). A simple Cauchy–Schwarz inequality $\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle \leq \|\boldsymbol{\eta}\|_2 \cdot \|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\|_2$ is insufficient to provide a tight result since $\|\boldsymbol{\eta}\|_2$ scales as $\frac{1}{\sqrt{M}}$ as discussed after (34). Instead, we exploit the randomness of $\boldsymbol{\eta}$ and use the following concentration bound for multinomial random variables, which is proved in Lemma 14 of Appendix F and is derived based on [103].

Lemma 3. *Suppose* $\{(f_{i,k},\dots,f_{i,K})\}, i=1,\dots,Q$ *are mutually independent and follow the multinomial distribution* $\mathrm{Multinomial}(M,\boldsymbol{p}_i)$ *where* $\sum_{k=1}^K f_{i,k}=M$ *and* $\boldsymbol{p}_i=[p_{i,1},\cdots,p_{i,K}]$. *Let* $a_{i,1},\dots,a_{i,K}$ *be fixed. Then, for any* $t>0$*,*

$$\mathbb{P}\left(\sum_{i=1}^{Q}\sum_{k=1}^{K}a_{i,k}(\frac{f_{i,k}}{M}-p_{i,k})>t\right)\leq e^{-\frac{Mt}{4a_{\max}}\min\left\{1,\frac{a_{\max}t}{4\sum_{i=1}^{Q}\sum_{k=1}^{K}a_{i,k}^2p_{i,k}}\right\}}+e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{K}a_{i,k}^2p_{i,k}}}, \tag{40}$$

where $a_{\max}=\max_{i,k}|a_{i,k}|$.

One may not be able to directly apply the above result for $\langle\boldsymbol{\eta},\mathcal{A}^Q(\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star)\rangle$ since $\widehat{\boldsymbol{\rho}}$ depends on $\boldsymbol{\eta}$. We address this issue by using the covering argument to bound $\langle\boldsymbol{\eta},\mathcal{A}^Q(\boldsymbol{\rho}-\boldsymbol{\rho}^\star)\rangle$ for all possible $\boldsymbol{\rho}$. We refer to Appendix E for the detailed analysis. We now summarize the main result as follows.

Theorem 5. *Given an MPO state* $\boldsymbol{\rho}^\star\in\mathbb{C}^{d^n\times d^n}$ *of the form* (7) *with MPO ranks* $\boldsymbol{r}$*, independently generate* Q *Haar-distributed random unitary matrices* $\begin{bmatrix}\boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n}\end{bmatrix}, i=1,\dots,Q$. *Use each induced rank-one POVM* $\{\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H\}_{k=1}^{d^n}$ *to measure the state* M *times and get the empirical measurements* $\widehat{\boldsymbol{p}}_i$. *For any* $\epsilon>0$*, suppose* $Q\gtrsim nd^2\overline{r}^2(\log n)$ *and*

$$QM\gtrsim\frac{nd^2\overline{r}^2\log n(\log Q+n\log d)^2}{\epsilon^2},\quad \overline{r}=\max_{i=1,\dots n-1}r_i. \tag{41}$$

Then any global solution $\widehat{\boldsymbol{\rho}}$ *of* (33) *satisfies*

$$\|\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star\|_F\leq\epsilon \tag{42}$$

with probability at least $\min\{1-e^{-\alpha_2Q},1-e^{-\alpha_3(\log Q+n\log d)}-e^{-\alpha_4nd^2\overline{r}^2\log n}\}$*, where* α_3 *and* α_4 *are positive constants,* α_2 *corresponds to constants of the probability in Theorem 4.*

Theorem 5 ensures a stable recovery of the ground-truth state when the total number of state copies QM only grows polynomially (n^3) in terms of the number of qudits n, as the order specified in (41). If we ignore the $\log Q$ term, which exists due a to proof artifact and which we conjecture can be removed, then (41) only requires QM to be sufficiently large, without any requirement on the number of measuring times M for each POVM. In other words, Theorem 5 provides theoretical support for the practical use of single-shot measurements (i.e., $M=1$ where each POVM is measured only once) that are used in [32,43]. Note that the orders of the polynomial in (41), particularly in terms of n, are fairly large compared to the number $O(nd^2\overline{r}^2)$ of degrees of freedom of the MPO and may not be optimal. For this reason, we conjecture that the bound in (41) could be further improved, such as by removing the term $(\log Q+n\log d)^2$ that extends the bound beyond the number $O(nd^2\overline{r}^2)$ of degrees of freedom of the MPO. We refer to Section 6 for additional detailed discussion.

The requirement $Q\gtrsim nd^2\overline{r}^2\log n$ and failure probability $e^{-\alpha_2Q}$ are inherited from Theorem 4 for a stable embedding via the Q POVMs $\{\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H\}$. As discussed right after Theorem 4, we conjecture that Theorem 4 holds with $Q=1$ by setting $K=d^n$. If this is the case, then the requirement $Q\gtrsim nd^2\overline{r}^2\log n$ can also be dropped, and Theorem 5 would also hold for $Q=1$. In the next section, we will use experiments to demonstrate that a single POVM is sufficient to stably recover $\boldsymbol{\rho}^\star$.

5 Numerical Experiments

In this section, we perform numerical experiments on quantum state tomography for MPOs to illustrate our theoretical results. Due to computational constraints, we conduct experiments on real matrix product states (MPSs, which are pure states) of the form $\boldsymbol{\rho}^\star=\boldsymbol{u}^\star\boldsymbol{u}^{\star T}$, where $\boldsymbol{u}^\star\in\mathbb{R}^{d^n\times 1}$ satisfies $\|\boldsymbol{u}^\star\|_2=1$ and its $(i_1\cdots i_n)$-element can be represented in a matrix product form similar to the MPO form in (7):

$$\boldsymbol{u}^\star(i_1\cdots i_n)=\boldsymbol{U}_1^{\star i_1}\cdots\boldsymbol{U}_n^{\star i_n}.$$

Here, each matrix $\boldsymbol{U}_\ell^{\star i_\ell}$ has size $r\times r$, except for $\boldsymbol{U}_1^{\star i_1}$ and $\boldsymbol{U}_n^{\star i_n}$ that have dimension of $1\times r$ and $r\times 1$, respectively. We generate each MPS $\boldsymbol{u}^\star$ by first generating a random Gaussian vector of length d^n and then applying the sequential SVD [27] to truncate it to an MPS, which we finally normalize to have unit length. As a consequence, entry $\boldsymbol{\rho}^\star(i_1\cdots i_n, j_1\cdots j_n)$ can be expressed as

$$\begin{aligned}\boldsymbol{\rho}^\star(i_1\cdots i_n, j_1\cdots j_n) &= \boldsymbol{U}_1^{\star i_1}\cdots\boldsymbol{U}_n^{\star i_n}\boldsymbol{U}_1^{\star j_1}\cdots\boldsymbol{U}_n^{\star j_n}\\ &= (\boldsymbol{U}_1^{\star i_1}\cdots\boldsymbol{U}_n^{\star i_n})\otimes(\boldsymbol{U}_1^{\star j_1}\cdots\boldsymbol{U}_n^{\star j_n})\\ &= \underbrace{(\boldsymbol{U}_1^{\star i_1}\otimes\boldsymbol{U}_1^{\star j_1})}_{\boldsymbol{X}_1^{\star i_1,j_1}}\cdots\underbrace{(\boldsymbol{U}_n^{\star i_n}\otimes\boldsymbol{U}_n^{\star j_n})}_{\boldsymbol{X}_n^{\star i_n,j_n}},\end{aligned}$$

where $\otimes$ denotes the Kronecker product. Thus, $\boldsymbol{\rho}^\star = \boldsymbol{u}^\star\boldsymbol{u}^{\star T}$ is also an MPO with MPO ranks $r_1 = \cdots = r_{n-1} = r^2$.

To illustrate that Theorem 5 might hold even with $Q = 1$, we only use a single Haar random rank-one POVM in the experiments. We generate a real Haar-distributed random unitary matrix $\begin{bmatrix}\boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{d^n}\end{bmatrix} \in \mathbb{R}^{d^n\times d^n}$. Each population measurement (2) can then be rewritten as

$$p_k = \operatorname{trace}(\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\rho}^\star) = \left|\boldsymbol{\phi}_k^T\boldsymbol{u}^\star\right|^2.$$

This is our reason for considering a pure state $\boldsymbol{\rho}^\star$ as it reduces the complexity for computing $\operatorname{trace}(\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\rho}^\star)$ from $O(d^{2n})$ to $O(d^n)$.

We use the induced POVM to measure the state $\boldsymbol{\rho}^\star$ M times to get the empirical measurements $\widehat{\boldsymbol{p}}$. With the obtained measurements, as in (33), we attempt to recover the MPS $\boldsymbol{u}^\star$ (and hence $\boldsymbol{\rho}^\star$) by minimizing the following constrained mean squared error loss

$$\begin{aligned}&\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}\in\mathbb{U}_r}\frac{1}{2}\sum_{i=1}^{d^n}(|\boldsymbol{\phi}_i^T\boldsymbol{u}|^2 - \widehat{p}_i)^2,\\ &\mathbb{U}_r = \left\{\boldsymbol{u}\in\mathbb{R}^{d^n} : \boldsymbol{u}(i_1\cdots i_n) = {\boldsymbol{U}_1}^{i_1}\cdots{\boldsymbol{U}_n}^{i_n}, \boldsymbol{U}_1^{i_1}\in\mathbb{R}^{1\times r}, \boldsymbol{U}_n^{i_n}\in\mathbb{R}^{r\times 1}, \boldsymbol{U}_\ell^{i_\ell}\in\mathbb{R}^{r\times r}, 1 < i < n\right\},\end{aligned} \tag{43}$$

which has the same form as (33).

As in [49], we solve (43) by the following iterative hard thresholding (IHT, i.e., projected gradient descent):

$$\boldsymbol{u}_{t+1} = \mathcal{P}_{\mathbb{U}_r}\left(\boldsymbol{u}_t - \mu\sum_{i=1}^{d^n}(|\boldsymbol{\phi}_i^T\boldsymbol{u}_t|^2 - \widehat{p}_i)\boldsymbol{\phi}_i\boldsymbol{\phi}_i^T\boldsymbol{u}_t\right), \tag{44}$$

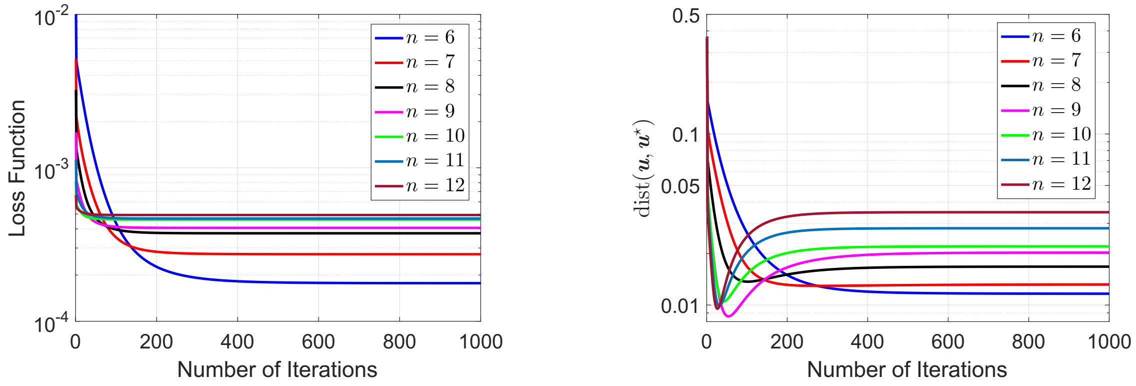
where μ is the step size and $\mathcal{P}_{\mathbb{U}_r}$ denotes the projection onto the MPS set $\mathbb{U}_r$, which can be approximately computed via a sequential SVD algorithm [27]. We adopt this approach for computing an approximate projection in the following experiments.

Since our goal is to verify how the global solution $\widehat{\boldsymbol{u}}$ behaves, to ensure the convergence to a global solution, we use a good initialization $\boldsymbol{u}_0 = \frac{\boldsymbol{u}^\star+\lambda\boldsymbol{v}}{\|\boldsymbol{u}^\star+\lambda\boldsymbol{v}\|_2}$ where $\boldsymbol{v}$ is randomly generated from the unit sphere of $\mathbb{R}^{d^n}$. In all the experiments, we set $\lambda = 0.7$ so that the initialization $\boldsymbol{u}_0$ is still not very close to the ground truth $\boldsymbol{u}^\star$. Since the gradient becomes exponentially small in n, which can be observed by using the same argument in (37) for $\boldsymbol{\phi}_i^T\boldsymbol{u}_t$, we set the step size $\mu = 0.01\times d^n$. The solution $\widehat{\boldsymbol{u}}$ is obtained by running the IHT algorithm (44) until convergence. Since the factorization $\boldsymbol{\rho}^\star = \boldsymbol{u}^\star\boldsymbol{u}^{\star T}$ is not unique as $\boldsymbol{\rho}^\star = (-\boldsymbol{u}^\star)(-\boldsymbol{u}^\star)^T$ also holds, we measure the quality of the recovered $\widehat{\boldsymbol{u}}$ by the following distance

$$\operatorname{dist}(\widehat{\boldsymbol{u}}, \boldsymbol{u}^\star) = \min\left\{\|\widehat{\boldsymbol{u}} - \boldsymbol{u}^\star\|_2^2, \|\widehat{\boldsymbol{u}} + \boldsymbol{u}^\star\|_2^2\right\}. \tag{45}$$

For each experiment, we conduct 10 Monte Carlo trials and take the average recovery distance over the 10 trials.

Experimental results We first set $M = 1000$, $r = 2$, and $d = 2$ and examine the convergence of the IHT algorithm defined in (44). Figure 3(a) shows the convergence of the algorithm in minimizing the loss function defined in (43); it can be observed that the IHT algorithm converges relatively fast. Figure 3(b) plots the learning curves in terms of the recovery error for the ground-truth $\boldsymbol{u}^\star$ as defined in (45). We first note that the initialization $\boldsymbol{u}_0$ is not close to the ground truth $\boldsymbol{u}^\star$, which is consistent with our choice of initialization described above. After convergence, the algorithm produces a much better recovery of $\boldsymbol{u}^\star$ and the recovery error increases steadily as n increases. It is also of



(a)

(b)

Figure 3: Illustration of convergence of IHT in (44) in terms of (a) loss function defined in (43), and (b) recovery error defined in (45) for different n with $M = 1000$, $r = 2$, and $d = 2$.

interest to note that when n increases while M remains the same, the recovery error curve exhibits a "U-shape" that first decreases, followed by an increasing trend. In other words, if the algorithm stops appropriately at the initial phase, it produces an iterate much closer to the ground truth than the final one. This is sometimes called algorithmic bias and can be exploited to produce a better solution [104–106]. But we highlight that we do not use this early-stopping approach here, and instead run the algorithm until convergence and use the final iterate since our goal is to verify the properties of the global minimizer.

Next, in Figure 4, we plot the recovery error as a function of n for various values of M and r. As expected, the recovery error increases when M decreases or r increases, but the IHT algorithm produces stable performance in all cases. We also observe that the recovery error increases only polynomially rather than exponentially with n, which is consistent with Theorem 5.

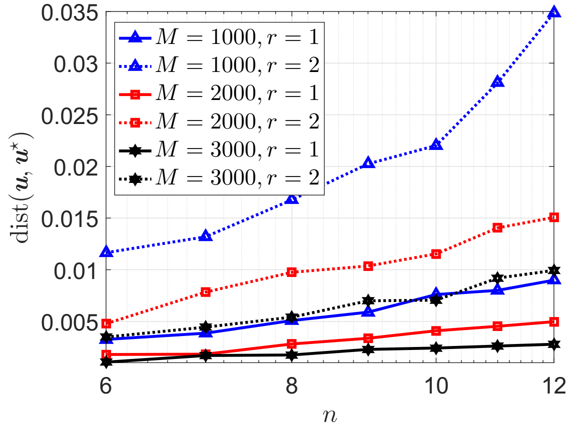


Figure 4: IHT recovery error as the number of qudits n increases with several choices of M and r.

6 Discussion and Conclusion

In this paper, we have studied sampling bounds for recovering structured quantum states that can be represented as matrix product operators (MPOs). We first established a non-asymptotic lower bound on the number of requisite measurements to ensure a stable embedding of MPOs under several choices of random measurement ensembles, including generic subgaussian measurements, rank-one Gaussian measurement ensembles, and Haar random rank-one POVMs. We then established theoretical bounds on the accuracy of a constrained least-squares estimator for recovering an MPO by using its empirical measurements obtained from multiple Haar random rank-one POVMs. Our research shows that a stable recovery guarantee requires only *polynomial* growth in the total number of state copies relative to the number of qudits. Thus, these results support the growing evidence for using MPOs for quantum state tomography and may have implications for the advancement of more efficient quantum state tomography methods in the future. Our findings suggest interesting directions for enhancing the current results or expanding our research to a more practical context. We elaborate on these possibilities below.

Stable embedding for MPOs with a single Haar random rank-one POVM As discussed right after Theorem 4, we conjecture that a single Haar random rank-one POVM is sufficient to establish stable embeddings for MPOs. This is supported by our numerical experiments with measurements from a single POVM to recover the MPO state. One possible approach is to exploit the fact that the local correlations between the columns in the unitary matrix are very weak because orthogonality is a global property [96]. Incorporating this property into Mendelson's small ball method presents a challenge, however. Another approach is to exploit the connection between the unitary matrix and the Gaussian distribution, as used in [95] for studying rank-one tight frame measurements.

Improving sampling complexity for the number of state copies In Theorem 5, we established a recovery guarantee for MPOs from Haar random rank-one POVM measurements. The result requires a total number of state copies $QM = \widetilde{\Omega}\big(\frac{n^3 d^2 \overline{r}^2}{\epsilon^2}\big)$. This sampling complexity is probably not optimal; one may compare it to $O(nd^2\overline{r}^2)$, the number of degrees of freedom in the MPOs. Below we consider rank-one Gaussian measurements and use an alternative approach to establish a recovery guarantee.

Consider the rank-one Gaussian measurement ensembles $\{\boldsymbol{a}_k\boldsymbol{a}_k^H\}_{k=1}^K$. The Chernoff bound [107] implies that for sufficiently large K, $\frac{1}{K}\sum_{k=1}^K \boldsymbol{a}_k\boldsymbol{a}_k^H \approx \boldsymbol{I}_{d^n}$. Hence, we may view $\{\frac{1}{K}\boldsymbol{a}_k\boldsymbol{a}_k^H\}_{k=1}^K$ as being similar to a POVM, though the rank-one measurement matrices do not exactly sum to the identity. Then, we may define the population measurements as $p_k = \langle \frac{1}{K}\boldsymbol{a}_k\boldsymbol{a}_k^H, \boldsymbol{\rho}^\star\rangle, k = 1,\dots,K$, and denote by $\mathcal{A}$ the associated linear measurement operator such that $\mathcal{A}(\boldsymbol{\rho}^\star) = \{p_k\}$. Also, the empirical measurements obtained by measuring the states M times are denoted by $\widehat{p}_k = f_k/M$, where $f_1,\dots,f_K$ follow the multinomial distribution $\operatorname{Multinomial}(M,\boldsymbol{p})$ with parameters M and $\boldsymbol{p} = \begin{bmatrix} p_1 & \cdots & p_K\end{bmatrix}^\top$. Denote by $\boldsymbol{\eta}$ the measurement errors with entries $\eta_k = \widehat{p}_k - p_k, k = 1,\dots,K$.

Suppose we solve the same problem (33) (with $\mathcal{A}^Q$ replaced by $\mathcal{A}$) and denote its global solution as $\widehat{\boldsymbol{\rho}}$. It follows that

$$\|\mathcal{A}(\widehat{\boldsymbol{\rho}}) - \mathcal{A}(\boldsymbol{\rho}^\star) - \boldsymbol{\eta}\|_2 \le \|\mathcal{A}(\boldsymbol{\rho}^\star) - \mathcal{A}(\boldsymbol{\rho}^\star) - \boldsymbol{\eta}\|_2 = \|\boldsymbol{\eta}\|_2.$$

On the other hand, $\|\mathcal{A}(\widehat{\boldsymbol{\rho}}) - \mathcal{A}(\boldsymbol{\rho}^\star) - \boldsymbol{\eta}\|_2 \ge \|\mathcal{A}(\widehat{\boldsymbol{\rho}}) - \mathcal{A}(\boldsymbol{\rho}^\star)\|_2 - \|\boldsymbol{\eta}\|_2$, which together with the above equation gives

$$\|\mathcal{A}(\widehat{\boldsymbol{\rho}}) - \mathcal{A}(\boldsymbol{\rho}^\star)\|_2 \le 2\|\boldsymbol{\eta}\|_2. \tag{46}$$

According to Theorem 3, the left-hand side can be further lower bounded by $\|\mathcal{A}(\widehat{\boldsymbol{\rho}}) - \mathcal{A}(\boldsymbol{\rho}^\star)\|_2 \gtrsim \|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F/\sqrt{K}$. Note that the scaling is different from (22), which is due to the scaling difference between the measurement operator $\mathcal{A}$ and the one defined in Theorem 3. On the other hand, since $\mathbb{E}\,\boldsymbol{\eta} = 0$ and $\mathbb{E}\,\|\boldsymbol{\eta}\|_2^2 = \frac{1}{M^2}\sum_{k=1}^K Mp_k(1-p_k) \le \frac{1}{M}$, we can use a concentration inequality such as Chebyshev's inequality to obtain $\|\boldsymbol{\eta}\|_2 \lesssim \frac{1}{\sqrt{M}}$ with high probability. Plugging these equations into (46) gives

$$\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F \lesssim \sqrt{\frac{K}{M}}. \tag{47}$$

The above derivation is commonly used in studying recovery accuracy for inverse problems. On one hand, since Theorem 3 only requires $K = \widetilde{\Omega}(nd^2\overline{r}^2)$, by taking $K = \Omega(nd^2\overline{r}^2\log n)$ in (47), we observe that $M = \widetilde{\Omega}(nd^2\overline{r}^2/\epsilon^2)$ is sufficient to ensure $\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F \le \epsilon$. As demonstrated in this context, this approach often leads to an optimal, or nearly optimal, recovery bound when using a minimal yet sufficient number of measurements. As another example, the work [23] employs this approach to establish a recovery bound for low-rank states. But on the other hand, recall that the above derivation is based on the assumption that $\frac{1}{K}\sum_{k=1}^K \boldsymbol{a}_k\boldsymbol{a}_k^H \approx \boldsymbol{I}_{d^n}$, which holds only when $K \gtrsim d^n$. If we plug this into (47), then $M \gtrsim d^n/\epsilon^2$ is required to ensure ϵ-accuracy. This also illustrates the challenge described in Section 4.1. Nevertheless, the discussion above suggests that it may still be possible to improve upon our current bound for the total number of state copies.

Convergence of IHT and other efficient optimization algorithms The algorithmic aspect is not the focus of this work. In our experiments, we employ an IHT algorithm, but we do not provide a formal guarantee for that algorithm. It will be of interest to develop a convergence guarantee for this algorithm. As discussed in [49], one potential challenge is to find a good initialization that allows IHT to converge quickly to the target solution. On the other hand, the IHT algorithm requires performing a sequential SVD algorithm in each iteration, which could be computationally expensive, especially for large quantum systems. Consequently, exploring alternative optimization algorithms that offer computational efficiency without the need for a projection step and can effectively handle an increasing number of qudits has become an area of great interest.

Extension to local measurements In this paper, we primarily focus on rank-one POVMs with Haar-distributed unitary matrices. Such measurements are known as global measurements since the unitary matrix will rotate the entire system of qudits simultaneously. This poses challenges in performing these measurements with practical quantum circuits. Therefore, an important future direction we will pursue is to study other measurement settings, such as the unitary t-design [37,108] or even local measurements [15,109] that can be conducted efficiently on current quantum computers. Such measurement settings may also reduce the cost for computing the gradient of the least squares loss (33). It is also interesting to design measurement operators that can improve efficiency in both performing experimental measurements and post-processing for estimating the state, which is often achieved by using certain iterative algorithms.

Acknowledgment

We acknowledge funding support from NSF Grants No. CCF-1839232, PHY-2112893, CCF-2106834 and CCF-2241298, as well as the W. M. Keck Foundation. We thank the Ohio Supercomputer Center for providing the computational resources needed in carrying out this work. Finally, we are grateful to Rungang Han, Holger Rauhut, Gongguo Tang, and Roman Vershynin for many valuable discussions and to Alireza Goldar for helpful comments on the manuscript.

Appendices

To simplify the notations, universal constants in each proof may share the same symbols (e.g., c_0), but they could represent different values.

A Proof of Theorem 2

A.1 Generic subgaussian measurements

We first extend the statement of Theorem 2 to generic subgaussian measurements and then prove this more general result.

Definition 3 (Subgaussian measurement ensembles [110])**.** *A complex random variable X is called L-subgaussian if there exists a constant $L > 0$ such that $\mathbb{E}\, e^{\mathscr{R}(tX)} \leq e^{L^2|t|^2/2}$ holds for all $t \in \mathbb{C}$. We say that $\mathcal{A} : \mathbb{C}^{d^n \times d^n} \to \mathbb{C}^K$ is an L-subgaussian measurement ensemble if all the elements of $\boldsymbol{A}_k, k = 1, \dots, K$ are independent L-subgaussian random variables with mean zero and variance one.*

Note that a complex-valued random variable X is subgaussian if and only if its both real part $\mathscr{R}(X)$ and imaginary part $\mathscr{I}(X)$ are real subgaussian random variables. We define the subgaussian norm of X as

$$\|X\|_{\psi_2} = \inf\left\{t > 0,\ \mathbb{E}\, e^{\frac{|X|^2}{t^2}} \leq 2\right\}. \tag{48}$$

Here are some classical examples of subgaussian distributions.

- (Gaussian) A standard complex Gaussian random variable $X = \mathscr{R}(X) + i\mathscr{I}(X)$ with $\mathscr{R}(X)$ and $\mathscr{I}(X)$ being independent and following $\mathcal{N}(0, \frac{1}{2})$, is a subgaussian random variable with $\|X\|_{\psi_2} \leq C$, where C is an absolute constant.
- (Bernoulli) A Bernoulli random variable X that takes values -1 and 1 with equal probability is a subgaussian random variable with $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$.

The following result establishes the RIP for an L-subgaussian measurement ensemble.

Theorem 6. *Suppose the linear map $\mathcal{A} : \mathbb{C}^{d^n \times d^n} \to \mathbb{C}^K$ is an L-subgaussian measurement ensemble defined in Definition 3. Then, with probability at least $1-\overline{\epsilon}$, $\mathcal{A}$ satisfies the $\delta_{\overline{r}}$-RIP as in* (16) *for MPOs given that*

$$K \geq C \cdot \frac{1}{\delta_{\overline{r}}^2} \cdot \max\left\{nd^2\overline{r}^2(\log n\overline{r}), \log(1/\overline{\epsilon})\right\}, \tag{49}$$

where C is a universal constant depending only on L.

A.2 Covering number for MPOs

To study the RIP or similar properties for MPOs, we need to first compute the covering number for the set of unit-norm MPOs. Since a unit-norm MPO can always be written in the canonical form, we consider the set $\overline{\mathbb{X}}_{\overline{r}}$ defined in (15). Note that the definition of (15) is different from (11) since in (15), we assume $\|\boldsymbol{\rho}\|_F = 1$ such that $\sum_{i_n,j_n} {\boldsymbol{X}_n^{i_n,j_n}}^H \boldsymbol{X}_n^{i_n,j_n} = 1$.

We say $\widetilde{\mathbb{X}}_{\overline{r}}$ is an ϵ-net of $\overline{\mathbb{X}}_{\overline{r}}$ if for any $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$, there exists $\overline{\boldsymbol{\rho}} \in \widetilde{\mathbb{X}}_{\overline{r}}$ such that $\|\boldsymbol{\rho} - \overline{\boldsymbol{\rho}}\|_F \leq \epsilon$. Here we use the Frobenius norm to quantify the distance, but one may use other metrics depending on the application. We now provide the covering number of $\widetilde{\mathbb{X}}_{\overline{r}}$.

Lemma 4. *There exists an ϵ-net $\widetilde{\mathbb{X}}_{\overline{r}}$ for $\overline{\mathbb{X}}_{\overline{r}}$ in* (11) *under the Frobenius norm obeying*

$$\left|\widetilde{\mathbb{X}}_{\overline{r}}\right| \leq \left(\frac{3n\overline{r}}{\epsilon}\right)^{nd^2\overline{r}^2}, \tag{50}$$

where $\left|\widetilde{\mathbb{X}}_{\overline{r}}\right|$ denotes the number of elements in the set $\widetilde{\mathbb{X}}_{\overline{r}}$.

Proof. Denote by

$$L(\boldsymbol{X}_\ell) = \begin{bmatrix} \boldsymbol{X}_\ell^{1,1} \\ \vdots \\ \boldsymbol{X}_\ell^{d,d} \end{bmatrix} \in \mathbb{C}^{d^2 r_{\ell-1} \times r_\ell}, \text{ which satisfies } L(\boldsymbol{X}_\ell)^H L(\boldsymbol{X}_\ell) = \sum_{i_\ell, j_\ell} {\boldsymbol{X}_\ell^{i_\ell,j_\ell}}^H \boldsymbol{X}_\ell^{i_\ell,j_\ell} = \mathbf{I}_{r_\ell}. \tag{51}$$

For covering $\mathbb{O}_{d^2r_{\ell-1},r_\ell} = \left\{L(\boldsymbol{X}_\ell) \in \mathbb{C}^{d^2r_{\ell-1}\times r_\ell}, L(\boldsymbol{X}_\ell)^H L(\boldsymbol{X}_\ell) = \mathbf{I}_{r_\ell}\right\}$, which contains matrices with unit-norm and orthogonal column vectors, it is beneficial to use the $\|\cdot\|_{1,2}$ norm which counts the largest energy of each column, i.e., $\|\boldsymbol{A}\|_{1,2} = \max_i \|\boldsymbol{A}(:,i)\|_2$. By relaxing $\mathbb{O}_{d^2r_{\ell-1},r_\ell}$ to the set of matrices with unit-norm vectors, the standard result on the covering number of unit ball implies that there exists an ϵ_ℓ-net $\overline{\mathbb{O}}_{d^2r_{\ell-1},r_\ell}$ for $\mathbb{O}_{d^2r_{\ell-1},r_\ell}$ obeying $|\overline{\mathbb{O}}_{d^2r_{\ell-1},r_\ell}| \leq \left(\frac{3}{\epsilon_\ell}\right)^{d^2r_{\ell-1}r_\ell} \leq \left(\frac{3}{\epsilon_\ell}\right)^{d^2\overline{r}^2}$. Then we define the set

$$\widetilde{\mathbb{X}}_{\overline{r}} := \{\overline{\boldsymbol{\rho}} : \overline{\boldsymbol{\rho}}(i_1\cdots i_n, j_1\cdots j_n) = \Pi_{\ell=1}^n \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell},\ L(\overline{\boldsymbol{X}}_\ell) = \begin{bmatrix} \overline{\boldsymbol{X}}_\ell^{1,1} \\ \vdots \\ \overline{\boldsymbol{X}}_\ell^{d,d} \end{bmatrix} \in \overline{\mathbb{O}}_{d^2r_{\ell-1},r_\ell},\ \forall \ell \in [n]\}, \tag{52}$$

which obeys

$$\left|\widetilde{\mathbb{X}}_{\overline{r}}\right| \leq \prod_\ell \left(\frac{3}{\epsilon_\ell}\right)^{d^2\overline{r}^2}. \tag{53}$$

We now verify that $\widetilde{\mathbb{X}}_{\overline{r}}$ is an ϵ-net for $\overline{\mathbb{X}}_{\overline{r}}$ by appropriately selecting ϵ_ℓ. For any $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$ with $\boldsymbol{\rho}(i_1 \dots i_n, j_1 \dots j_n) = \Pi_{\ell=1}^n \boldsymbol{X}_\ell^{i_\ell, j_\ell}$, we construct $\overline{\boldsymbol{\rho}}$ with $\overline{\boldsymbol{\rho}}(i_1 \dots i_n, j_1 \dots j_n) = \Pi_{\ell=1}^n \overline{\boldsymbol{X}}_\ell^{i_\ell, j_\ell}$ where $\left\| L(\boldsymbol{X}_\ell) - L(\overline{\boldsymbol{X}}_\ell) \right\|_{1,2} \le \epsilon_\ell$. Then we have

$$\begin{aligned}
\Gamma &= \|\boldsymbol{\rho} - \overline{\boldsymbol{\rho}}\|_F^2 = \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \boldsymbol{X}_1^{i_1,j_1} \boldsymbol{X}_2^{i_2,j_2} \cdots \boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_1^{i_1,j_1} \overline{\boldsymbol{X}}_2^{i_2,j_2} \cdots \overline{\boldsymbol{X}}_n^{i_n,j_n} \right|^2 \\
&= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \sum_{\ell=1}^n \left(\overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{\ell-1}^{i_{\ell-1},j_{\ell-1}} \boldsymbol{X}_\ell^{i_\ell,j_\ell} \cdots \boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell} \boldsymbol{X}_{\ell+1}^{i_{\ell+1},j_{\ell+1}} \cdots \boldsymbol{X}_n^{i_n,j_n} \right) \right|^2 \\
&\le n \sum_{\ell=1}^n \underbrace{\sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{\ell-1}^{i_{\ell-1},j_{\ell-1}} \boldsymbol{X}_\ell^{i_\ell,j_\ell} \cdots \boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell} \boldsymbol{X}_{\ell+1}^{i_{\ell+1},j_{\ell+1}} \cdots \boldsymbol{X}_n^{i_n,j_n} \right|^2}_{\Gamma_\ell} \\
&\le n \sum_{\ell=1}^n r_\ell \epsilon_\ell^2, \qquad (54)
\end{aligned}$$

where the last line uses the inequalities $\Gamma_n \le \epsilon_n^2$ and $\Gamma_\ell \le r_\ell \epsilon_\ell^2, \ell = 1, \dots, n-1$ which can be proved by

$$\begin{aligned}
\Gamma_n &= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}} \boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}} \overline{\boldsymbol{X}}_n^{i_n,j_n} \right|^2 \\
&= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n}) \right|^2 \\
&= \sum_{i_2,\dots,i_n,j_2,\dots,j_n} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n})^H \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}\,H} \cdots \underbrace{\sum_{i_1,j_1} (\overline{\boldsymbol{X}}_1^{i_1,j_1\,H} \overline{\boldsymbol{X}}_1^{i_1,j_1})}_{=\mathbf{I}_{r_1}} \cdots \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n}) \\
&= \sum_{i_2,\dots,i_n,j_2,\dots,j_n} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n})^H \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}\,H} \cdots \boldsymbol{X}_2^{i_2,j_2\,H} \boldsymbol{X}_2^{i_2,j_2} \cdots \overline{\boldsymbol{X}}_{n-1}^{i_{n-1},j_{n-1}} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n}) \\
&= \cdots = \sum_{i_n,j_n} (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n})^H (\boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_n^{i_n,j_n}) = \left\| L(\boldsymbol{X}_n) - L(\overline{\boldsymbol{X}}_n) \right\|_2^2 \le \epsilon_n^2, \qquad (55)
\end{aligned}$$

and similarly

$$\begin{aligned}
\Gamma_\ell &= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{\ell-1}^{i_{\ell-1},j_{\ell-1}} \boldsymbol{X}_\ell^{i_\ell,j_\ell} \cdots \boldsymbol{X}_n^{i_n,j_n} - \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{\ell-1}^{i_{\ell-1},j_{\ell-1}} \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell} \cdots \boldsymbol{X}_n^{i_n,j_n} \right|^2 \\
&= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \left| \overline{\boldsymbol{X}}_1^{i_1,j_1} \cdots \overline{\boldsymbol{X}}_{\ell-1}^{i_{\ell-1},j_{\ell-1}} (\boldsymbol{X}_\ell^{i_\ell,j_\ell} - \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell}) \cdots \boldsymbol{X}_n^{i_n,j_n} \right|^2 \\
&= \sum_{i_1,\dots,i_n,j_1,\dots,j_n} \boldsymbol{X}_n^{i_n,j_n\,H} \cdots (\boldsymbol{X}_\ell^{i_\ell,j_\ell} - \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell})^H \cdots \underbrace{\sum_{i_1,j_1} (\overline{\boldsymbol{X}}_1^{i_1,j_1\,H} \overline{\boldsymbol{X}}_1^{i_1,j_1})}_{=\mathbf{I}_{r_1}} \cdots (\boldsymbol{X}_\ell^{i_\ell,j_\ell} - \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell}) \cdots \boldsymbol{X}_n^{i_n,j_n} \\
&= \cdots = \sum_{i_\ell,\dots,i_n,j_\ell,\dots,j_n} \boldsymbol{X}_n^{i_n,j_n\,H} \cdots (\boldsymbol{X}_\ell^{i_\ell,j_\ell} - \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell})^H (\boldsymbol{X}_\ell^{i_\ell,j_\ell} - \overline{\boldsymbol{X}}_\ell^{i_\ell,j_\ell}) \cdots \boldsymbol{X}_n^{i_n,j_n} \\
&\le \left\| L(\boldsymbol{X}_\ell) - L(\overline{\boldsymbol{X}}_\ell) \right\|_F^2 \underbrace{\sum_{i_{\ell+1},\dots,i_n,j_{\ell+1},\dots,j_n} \boldsymbol{X}_n^{i_n,j_n\,H} \cdots \boldsymbol{X}_{\ell+1}^{i_{\ell+1},j_{\ell+1}\,H} \boldsymbol{X}_{\ell+1}^{i_{\ell+1},j_{\ell+1}} \cdots \boldsymbol{X}_n^{i_n,j_n}}_{=1} \le r_\ell \epsilon_\ell^2
\end{aligned}$$

for all $\ell \le n-1$. Therefore, we can choose $\epsilon_\ell = \frac{\epsilon}{\overline{r}n}$ in (53) to ensure $\widetilde{\mathbb{X}}_{\overline{r}}$ as an ϵ-net for $\overline{\mathbb{X}}_{\overline{r}}$ (as $\|\boldsymbol{\rho} - \overline{\boldsymbol{\rho}}\|_F^2 \le$

$n\sum_{\ell=1}^n r_\ell \epsilon_\ell^2 \le \epsilon^2$) and such $\widetilde{\mathbb{X}}_{\overline{r}}$ obeys

$$\left|\widetilde{\mathbb{X}}_{\overline{r}}\right| \le \left(\frac{3n\overline{r}}{\epsilon}\right)^{nd^2\overline{r}^2}. \tag{56}$$

This completes the proof of Lemma 4. □

A.3 Proof of Theorem 6

Using the covering number established in Lemma 4, we can now follow the arguments in [71] to establish the RIP for MPOs $\boldsymbol{\rho}$ under subgaussian measurements. Because of the linearity of the measurement operator $\mathcal{A}$, we note that there exists a complex-valued matrix $\boldsymbol{A}$ of size $K \times d^{2n}$ such that

$$\mathcal{A}(\boldsymbol{\rho}) = \boldsymbol{A}\,\text{vec}(\boldsymbol{\rho}), \tag{57}$$

where $\text{vec}(\boldsymbol{\rho}) \in \mathbb{C}^{d^{2n}}$ denotes the vectorization (in any predetermined order) of the MPO format $\boldsymbol{\rho}$. Note that our goal is to study the quantity $\frac{1}{K}\|\mathcal{A}(\boldsymbol{\rho})\|_2^2 = \|\frac{1}{\sqrt{K}}\mathcal{A}(\boldsymbol{\rho})\|_2^2$. Equivalently, there exists a vector $\boldsymbol{\xi} \in \mathbb{C}^{Kd^{2n}}$ (containing the row-wise vectorization of $\boldsymbol{A}$) such that

$$\frac{1}{\sqrt{K}}\mathcal{A}(\boldsymbol{\rho}) = \boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}, \tag{58}$$

where $\boldsymbol{V}_{\boldsymbol{\rho}}$ is an $K \times Kd^{2n}$ matrix given by

$$\boldsymbol{V}_{\boldsymbol{\rho}} = \frac{1}{\sqrt{K}}\begin{bmatrix} \text{vec}(\boldsymbol{\rho})^H & & & \\ & \text{vec}(\boldsymbol{\rho})^H & & \\ & & \ddots & \\ & & & \text{vec}(\boldsymbol{\rho})^H \end{bmatrix}. \tag{59}$$

Now we begin to prove Theorem 6.

Proof. For any $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$ in (11), we recall (16). Because $\boldsymbol{\xi}$ is a random vector, $\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}$ is also a random vector. We can compute the expectation of the energy of this random vector:

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 &= \mathbb{E}(\boldsymbol{\xi}^H\boldsymbol{V}_{\boldsymbol{\rho}}^H\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}) = \mathbb{E}\,\text{trace}(\boldsymbol{\xi}^H\boldsymbol{V}_{\boldsymbol{\rho}}^H\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}) = \mathbb{E}\,\text{trace}(\boldsymbol{V}_{\boldsymbol{\rho}}^H\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\boldsymbol{\xi}^H)\\ &= \text{trace}(\boldsymbol{V}_{\boldsymbol{\rho}}^H\boldsymbol{V}_{\boldsymbol{\rho}}\,\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\xi}^H)) = \text{trace}(\boldsymbol{V}_{\boldsymbol{\rho}}^H\boldsymbol{V}_{\boldsymbol{\rho}}\mathbf{I}) = \|\boldsymbol{\rho}\|_F^2.\end{aligned} \tag{60}$$

Here we used the fact that $\mathbb{E}\,\boldsymbol{\xi}\boldsymbol{\xi}^H = \mathbf{I}$ since, by assumption, all elements of $\boldsymbol{\xi}$ are independent mean-zero, variance one, L-subgaussian variables. Using (58), and (60), we note that proving that $\mathcal{A}$ satisfies the $\delta_{\overline{r}}$-RIP is equivalent to proving

$$\sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \left|\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2\right| \le \delta_{\overline{r}}. \tag{61}$$

We can view $\left|\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2\right|$ as a random process indexed by the variable $\boldsymbol{\rho}$, and our goal is to bound the supremum of this random process over the set $\overline{\mathbb{X}}_{\overline{r}}$. [71, Theorem 3] gives a mechanism to bound this supremum. Specifically, let $\mathcal{B} := \{\boldsymbol{V}_{\boldsymbol{\rho}} :\ \boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}\}$ and note that

$$\sup_{\boldsymbol{B}\in\mathcal{B}} \left|\|\boldsymbol{B}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{B}\boldsymbol{\xi}\|_2^2\right| = \sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \left|\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2\right|. \tag{62}$$

[71, Theorem 3] states that there exist constants c_1, c_2 (depending on L) such that for $t > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{B}\in\mathcal{B}} \left|\|\boldsymbol{B}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{B}\boldsymbol{\xi}\|_2^2\right| \ge c_1E + t\right) \le 2e^{-c_2\min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}}, \tag{63}$$

where E, U, and V are quantities defined as

$$\begin{aligned}
&E := \gamma_2(\mathcal{B}, \|\cdot\|_{2\to2})\left(\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) + d_F(\mathcal{B})\right) + d_F(\mathcal{B})d_{2\to2}(\mathcal{B}),\\
&V := d_4^2(\mathcal{B}),\\
&U := d_{2\to2}^2(\mathcal{B}),
\end{aligned} \tag{64}$$

and $d_F(\mathcal{B})$, $d_{2\to2}(\mathcal{B})$, $d_4^2(\mathcal{B})$, and $\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2})$ are quantities that we define and bound in the next paragraph.

In this paragraph, we bound the quantities E, U, and V appearing in (63). To do this, we define and bound $d_F(\mathcal{B})$, $d_{2\to2}(\mathcal{B})$, $d_4^2(\mathcal{B})$, and $\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2})$ which appear in the definitions of E, U, and V in (64). First,

$$d_F(\mathcal{B}) := \sup_{\boldsymbol{B}\in\mathcal{B}} \|\boldsymbol{B}\|_F^2 = \sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \|\boldsymbol{\rho}\|_F^2 = 1, \tag{65}$$

since every MPO format $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$ has unit norm. Second,

$$d_{2\to2}(\mathcal{B}) := \sup_{\boldsymbol{B}\in\mathcal{B}} \|\boldsymbol{B}\|_{2\to2} = \sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{\sqrt{K}}\|\boldsymbol{\rho}\|_F^2 = \frac{1}{\sqrt{K}}, \tag{66}$$

due to the block diagonal structure of $\boldsymbol{V}_{\boldsymbol{\rho}}$ (see (59)) and the normalization of all $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$. Third,

$$d_4(\mathcal{B}) := \sup_{\boldsymbol{B}\in\mathcal{B}} \left(\operatorname{trace}(\boldsymbol{B}^H\boldsymbol{B})^2\right)^{1/4} = K^{-1/4}; \tag{67}$$

see [71, Eqn. (65)] for an analogous derivation. Fourth,

$$\begin{aligned}
\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) &\le C\int_0^{d_{2\to2}(\mathcal{B})} \sqrt{\log\mathcal{N}(\mathcal{B}, \|\cdot\|_{2\to2}, u)}\, du\\
&= C\int_0^{\frac{1}{\sqrt{K}}} \sqrt{\log\mathcal{N}(\mathcal{B}, \|\cdot\|_{2\to2}, u)}\, du,
\end{aligned} \tag{68}$$

where the covering number $\mathcal{N}(\mathcal{B}, \|\cdot\|_{2\to2}, u)$ denotes the minimum cardinality of a u-net for $\mathcal{B}$ with respect to the norm $\|\cdot\|_{2\to2}$. As suggested by (66), the $\|\cdot\|_{2\to2}$ distance on $\mathcal{B}$ is equivalent to $\frac{1}{\sqrt{K}}$ times the squared Frobenius distance on $\overline{\mathbb{X}}_{\overline{r}}$. Therefore,

$$\mathcal{N}(\mathcal{B}, \|\cdot\|_{2\to2}, u) = \mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \frac{1}{\sqrt{K}}\|\cdot\|_F^2, u) = \mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \|\cdot\|_F, K^{1/4}\sqrt{u}). \tag{69}$$

Changing variables by letting $\epsilon = K^{1/4}\sqrt{u}$, (68) becomes

$$\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) \le 2C\frac{1}{\sqrt{K}}\int_0^1 \epsilon\sqrt{\log\mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \|\cdot\|_F, \epsilon)}\, d\epsilon \le C\frac{1}{\sqrt{K}}\int_0^1 \sqrt{\log\mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \|\cdot\|_F, \epsilon)}\, d\epsilon, \tag{70}$$

where the factor of 2 has been absorbed into the universal constant C. Now, by directly applying Lemma 4, we have that

$$\mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \|\cdot\|_F, \epsilon) \le \left(\frac{3n\overline{r}}{\epsilon}\right)^{nd^2\overline{r}^2}.$$

Therefore,

$$\begin{aligned}\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) &\le C\frac{1}{\sqrt{K}}\int_0^1 \sqrt{\log \mathcal{N}(\overline{\mathbb{X}}_{\overline{r}}, \|\cdot\|_F, \epsilon)}\, d\epsilon \\ &\le C\frac{1}{\sqrt{K}}\int_0^1 \sqrt{\log\left(\frac{3n\overline{r}}{\epsilon}\right)^{nd^2\overline{r}^2}}\, d\epsilon \\ &\le C\frac{1}{\sqrt{K}}\int_0^1 \sqrt{nd^2\overline{r}^2\log\left(\frac{3n\overline{r}}{\epsilon}\right)}\, d\epsilon \\ &\le C\sqrt{\frac{nd^2\overline{r}^2}{K}}\int_0^1 \sqrt{\log\left(\frac{3n\overline{r}}{\epsilon}\right)}\, d\epsilon \\ &\le C\sqrt{\frac{nd^2\overline{r}^2\log n\overline{r}}{K}}, \end{aligned} \tag{71}$$

where the last line follows from the fact that

$$\int_0^1 \sqrt{\log\left(\frac{3n\overline{r}}{\epsilon}\right)}\, d\epsilon \le C + \sqrt{\log(3n\overline{r})} \le C\sqrt{\log n\overline{r}},$$

and each appearance of C denotes an unspecified universal constant that may change from instance to instance. Putting together the above quantities, we have the following three numbers which appear in (63):

$$\begin{aligned} E &:= \gamma_2(\mathcal{B}, \|\cdot\|_{2\to2})\left(\gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) + d_F(\mathcal{B})\right) + d_F(\mathcal{B})d_{2\to2}(\mathcal{B}) \\ &= \gamma_2^2(\mathcal{B}, \|\cdot\|_{2\to2}) + \gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) + \frac{1}{\sqrt{K}}, \\ V &:= d_4^2(\mathcal{B}) = \frac{1}{\sqrt{K}}, \\ U &:= d_{2\to2}^2(\mathcal{B}) = \frac{1}{K}. \end{aligned} \tag{72}$$

Plugging (62) and (72) into (63), we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \left|\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2\right| \ge c_1(\gamma_2^2(\mathcal{B}, \|\cdot\|_{2\to2}) + \gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) + \frac{1}{\sqrt{K}}) + t\right) \le 2e^{-c_2\min\left\{Kt^2, Kt\right\}}. \tag{73}$$

Our goal is to find a value of K such that (61) holds with probability at least $1-\overline{\epsilon}$.

Let $t = \delta/2$ and recall that $\delta < 1$, so $\min\left\{Kt^2, Kt\right\} = K\delta^2/4$. If we choose $K > C\delta^{-2}\log(1/\overline{\epsilon})$ for an appropriately chosen constant C, we have $2e^{-c_2\min\left\{Kt^2, Kt\right\}} \le \overline{\epsilon}$. Next, using the bound on $\gamma_2^2(\mathcal{B}, \|\cdot\|_{2\to2})$ from (71), we see that by choosing

$$K \ge C\cdot\frac{nd^2\overline{r}^2(\log n\overline{r})}{\delta^2}, \tag{74}$$

for an appropriately chosen constant C, then we guarantee that

$$c_1(\gamma_2^2(\mathcal{B}, \|\cdot\|_{2\to2}) + \gamma_2(\mathcal{B}, \|\cdot\|_{2\to2}) + \frac{1}{\sqrt{K}}) \le \frac{\delta}{2}. \tag{75}$$

Putting all of the pieces together, we conclude that when (17) is satisfied, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \left|\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\,\|\boldsymbol{V}_{\boldsymbol{\rho}}\boldsymbol{\xi}\|_2^2\right| \ge \delta\right) \le \overline{\epsilon}. \tag{76}$$

We have thus proved that (61) holds with probability at least $1-\overline{\epsilon}$. This completes the proof of Theorem 6. □

B Proof of Theorem 3

Proof. In this section, we will apply Mendelson's small ball method to derive Theorem 3. According to Lemma 1, and supposing that $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K\}$ are selected independently from the standard normal distribution $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d^n})$, we need to bound

$$H_\xi(\overline{\mathbb{X}}_{\overline{r}}) = \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \mathbb{P}\{|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle| \geq \xi\} \tag{77}$$

and

$$W(\overline{\mathbb{X}}_{\overline{r}}) = \mathbb{E} \sup_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho}\rangle, \tag{78}$$

where $\{\epsilon_k\}$ is a Rademacher sequence independent from everything else.

- Lower bound of $H_\xi(\overline{\mathbb{X}}_{\overline{r}})$: To bound $H_\xi(\overline{\mathbb{X}}_{\overline{r}})$, we use the Paley-Zygmund inequality (Lemma 5). Specifically, we can get

$$\begin{aligned} H_\xi(\overline{\mathbb{X}}_{\overline{r}}) &= \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \mathbb{P}\left(|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle| \geq \xi\right) \\ &= \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \mathbb{P}\left(|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2 \geq \xi^2\right) \\ &\geq \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \mathbb{P}\left(|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2 \geq \frac{1}{2}\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2]\right) \\ &\geq \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \frac{(\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2])^2}{4\,\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^4]}, \ \forall \xi \leq \sqrt{\frac{1}{2}\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2]}, \end{aligned} \tag{79}$$

  where the first inequality follows because $\mathbb{P}\left(|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2 \geq \xi^2\right)$ is a decreasing function with respect to ξ, and the second inequality uses Lemma 5.

  Next we start to analyze $\frac{(\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2])^2}{4\,\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^4]}$. By the fact that $\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle$ is a second-order polynomial in the entries of Gaussian random vector $\boldsymbol{a}$, we can obtain $\|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle\|_{\psi_1} \leq c\|\boldsymbol{a}\|_{\psi_2}^2 \|\boldsymbol{\rho}\|_F \leq O(1)$ [111] for some constant c; thus, $\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle$ is a subexponential random variable. Hence, there exists α such that $\mathbb{E}\, e^{\alpha|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|}$ is finite. It follows from Lemma 6 that there exists a constant C_0 such that

$$\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^4] \leq C_0 \left(\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2]\right)^2. \tag{80}$$

  We need obtain ξ to finish the analysis. To that end, we bound the expectation $\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2]$. Since $\boldsymbol{\rho}$ is Hermitian, it has the eigenvalue decomposition $\boldsymbol{\rho} = \sum_{i=1}^{d^n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^H$. Using the same argument as in [19], we can obtain that

$$\mathbb{E}[|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2] \geq 1. \tag{81}$$

  Thus, we can set $\xi = \frac{1}{2}$. There exists a universal constant c_0 such that

$$\mathbb{P}\left(|\langle \boldsymbol{a}\boldsymbol{a}^H, \boldsymbol{\rho}\rangle|^2 \geq \frac{1}{2}\right) \geq c_0, \ \forall \boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}} \quad \Rightarrow \quad H_\xi(\overline{\mathbb{X}}_{\overline{r}}) \geq c_0. \tag{82}$$

- Upper bound of $W(\overline{\mathbb{X}}_{\overline{r}})$: As discussed above, $\langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho}\rangle, k = 1, \ldots, K$ are independent subexponential random variables since $\|\langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho}\rangle\|_{\psi_1} \leq c_1\|\boldsymbol{\rho}\|_F \|\boldsymbol{a}_k\|_{\psi_2}^2 \leq c_2$ [111] where c_1, c_2 are some universal constants and the second inequality follows from $\|\boldsymbol{\rho}\|_F = 1$ and $\|\boldsymbol{a}_k\|_{\psi_2} \leq O(1)$. In addition, we have $\mathbb{E}\langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho}\rangle = 0$ because of the Rademacher random variables ϵ_k.

By the analysis in the covering argument of Appendix B.1, when $K = \Omega(nd^2\overline{r}^2 \log n)$, we have

$$W(\overline{\mathbb{X}}_{\overline{r}}) = \mathbb{E} \sup_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho} \rangle \leq c_3 d \overline{r} \sqrt{n \log n}, \tag{83}$$

where c_3 is a positive constant.

- Contraction: Combining (82) and (83), we set $t = \frac{c_0\sqrt{K}}{2}$ and $K \geq \frac{256 c_3^2 n d^2 \overline{r}^2 \log n}{c_0^2}$ in (20), then get

$$\begin{aligned} \inf_{\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}} \left(\sum_{k=1}^{K} |\langle \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho} \rangle|^2 \right)^{\frac{1}{2}} &\geq \xi \sqrt{K} H_\xi(\overline{\mathbb{X}}_{\overline{r}}) - 2W(\overline{\mathbb{X}}_{\overline{r}}) - t\xi \\ &\geq \frac{c_0\sqrt{K}}{2} - 2c_3 d\overline{r}\sqrt{n \log n} - \frac{t}{2} \geq \frac{c_0\sqrt{K}}{8}. \end{aligned} \tag{84}$$

with probability $1 - e^{-\frac{c_0 K}{8}}$.

This completes the proof of Theorem 3. □

B.1 Proof of the upper bound for $W(\overline{\mathbb{X}}_{\overline{r}})$ in (83)

Proof. In this section, we apply a covering argument to prove (83). For an MPO $\boldsymbol{\rho}$ of the form (7), for simplicity, we denote it by $\boldsymbol{\rho} = [\boldsymbol{X}_1, \dots, \boldsymbol{X}_n] \in \overline{\mathbb{X}}_{\overline{r}}$. Also denote $\boldsymbol{A}_k = \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H$. For each set of matrices $\{L(\boldsymbol{X}_\ell) \in \mathbb{R}^{d^2 r_{\ell-1} \times r_\ell} : \|L(\boldsymbol{X}_\ell)\| \leq 1\}$ ($r_0 = 1$), according to [112], we can construct an ϵ-net $\{L(\boldsymbol{X}_\ell^{(1)}), \dots, L(\boldsymbol{X}_\ell^{(N_\ell)})\}$ with the covering number $N_\ell \leq (\frac{4+\epsilon}{\epsilon})^{d^2 r_{\ell-1} r_\ell}$ such that

$$\sup_{L(\boldsymbol{X}_\ell): \|L(\boldsymbol{X}_\ell)\| \leq 1} \min_{p_\ell \leq N_\ell} \|L(\boldsymbol{X}_\ell) - L(\boldsymbol{X}_\ell^{(p_\ell)})\| \leq \epsilon, \tag{85}$$

for all $\ell = 1, \dots, n-1$. Also, we can construct an ϵ-net $\{L(\boldsymbol{X}_n^{(1)}), \dots, L(\boldsymbol{X}_n^{(N_n)})\}$ for $\{L(\boldsymbol{X}_n) \in \mathbb{R}^{d^2 r_{n-1} \times 1} : \|L(\boldsymbol{X}_n)\|_F \leq 1\}$ such that

$$\sup_{L(\boldsymbol{X}_n): \|L(\boldsymbol{X}_n)\|_F \leq 1} \min_{p_n \leq N_n} \|L(\boldsymbol{X}_n) - L(\boldsymbol{X}_n^{(p_n)})\|_F \leq \epsilon, \tag{86}$$

with the covering number $N_n \leq (\frac{2+\epsilon}{\epsilon})^{d^2 r_{n-1}}$. Note that different from Lemma 4 that uses the $\|\cdot\|_{1,2}$ norm, here we use the spectral norm $\|\cdot\|$ and Frobenius norm $\|\cdot\|_F$ to define the covering numbers.

For simplicity, we use $\mathcal{I}$ to denote the index set $[N_1] \times \cdots \times [N_n]$. Denote by

$$[\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_n^\star] := \operatorname*{arg\,max}_{\substack{L(\boldsymbol{X}_\ell) \in \mathbb{R}^{d^2 r_{\ell-1} \times r_\ell} \\ \|\boldsymbol{X}_\ell\| \leq 1, \ell = 1, \dots, n-1 \\ \|\boldsymbol{X}_n\|_F \leq 1}} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \langle \boldsymbol{A}_k, [\boldsymbol{X}_1, \dots, \boldsymbol{X}_n] \rangle, \tag{87}$$

$$T := \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \langle \boldsymbol{A}_k, [\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_N^\star] \rangle. \tag{88}$$

According to the construction of the ϵ-nets, there exists $p = (p_1, \dots, p_n) \in \mathcal{I}$ such that

$$\|L(\boldsymbol{X}_\ell^\star) - L(\boldsymbol{X}_\ell^{(p_\ell)})\| \leq \epsilon, \ \ \ell = 1, \dots, n-1 \ \text{ and } \ \|L(\boldsymbol{X}_n^\star) - L(\boldsymbol{X}_n^{(p_n)})\|_F \leq \epsilon. \tag{89}$$

Now taking $\epsilon = \frac{1}{2n}$ gives

$$\begin{aligned}
T &= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle + \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_n^\star] - [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\right\rangle \\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle + \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left\langle \boldsymbol{A}_k, \sum_{a_1=1}^{n}[\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_{a_1-1}^{(p_1)}, \boldsymbol{X}_{a_1}^{(p_{a_1})} - \boldsymbol{X}_{a_1}^\star, \boldsymbol{X}_{a_1+1}^\star, \dots, \boldsymbol{X}_n^\star]\right\rangle \\
&\le \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle + n\epsilon T \\
&= \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle + \frac{T}{2}, \qquad (90)
\end{aligned}$$

where we write $[\boldsymbol{X}_1^\star, \dots, \boldsymbol{X}_n^\star] - [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]$ in the second line as the sum of n terms according to Lemma 10.

Notice that for any $\{L(\boldsymbol{X}_\ell)\}_{\ell \le n-1}$ and $L(\boldsymbol{X}_n)$, where $\|L(\boldsymbol{X}_\ell)\| \le 1$ and $\|L(\boldsymbol{X}_n)\|_F \le 1$, we have $\|\boldsymbol{\rho}\|_F = \|[\boldsymbol{X}_1, \dots, \boldsymbol{X}_n]\|_F \le 1$. As in the discussion in Appendix B, $\langle \boldsymbol{A}_k, \boldsymbol{\rho}\rangle = \langle \epsilon_k \boldsymbol{a}_k \boldsymbol{a}_k^H, \boldsymbol{\rho}\rangle$ is a centered subexponential random variable with subexponential norm of order $O(1)$, so we can use Lemma 7 to get

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle\right| \ge t\right) \le e^{1 - c_1 \min\{\frac{t^2}{c_2^2}, \frac{t\sqrt{K}}{c_2}\}}, \qquad (91)$$

where c_1 and c_2 are constants.

Combining (90) and (91) together yields

$$\begin{aligned}
\mathbb{P}(T \ge t) &\le \mathbb{P}\left(\max_{p_1, \dots, p_n}\left|\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\langle \boldsymbol{A}_k, [\boldsymbol{X}_1^{(p_1)}, \dots, \boldsymbol{X}_n^{(p_n)}]\rangle\right| \ge \frac{t}{2}\right) \\
&\le \left(\prod_{i=1}^{n} N_i\right) e^{1 - c_1 \min\{\frac{t^2}{c_2^2}, \frac{t\sqrt{K}}{c_2}\}} \\
&\le \left(\frac{4+\epsilon}{\epsilon}\right)^{d^2 r_1 + \sum_{i=2}^{n-1} d^2 r_{i-1} r_i + d^2 r_{n-1}} e^{1 - c_1 \min\{\frac{t^2}{c_2^2}, \frac{t\sqrt{K}}{c_2}\}} \\
&\le e^{1 - c_1 \min\{\frac{t^2}{c_2^2}, \frac{t\sqrt{K}}{c_2}\} + Cnd^2\overline{r}^2 \log n},
\end{aligned}$$

where $\overline{r} = \max_{i=1,\dots,n-1} r_i$, C is a universal constant, and the last line uses $\frac{4+\epsilon}{\epsilon} = \frac{4+\frac{1}{2n}}{\frac{1}{2n}} = 8n+1$ based on the assumption $\epsilon = \frac{1}{2n}$ in (90). Now choosing $K = c_3 n d^2 \overline{r}^2 \log n$ with a positive constant c_3 and plugging this into the above equation, we can find constants c_4 and c_5 such that

$$\mathbb{P}(T \ge t) \le e^{-c_4 t \sqrt{nd^2\overline{r}^2 \log n}}, \ \forall\, t \ge c_5\sqrt{nd^2\overline{r}^2 \log n},$$

which further implies that

$$\begin{aligned}
W(\overline{\mathbb{X}}_{\overline{r}}) &= \mathbb{E}\, T \\
&\le c_5\sqrt{nd^2\overline{r}^2 \log n} + \int_{c_5\sqrt{nd^2\overline{r}^2 \log n}}^{\infty} \mathbb{P}(T \ge t)\, dt \\
&\le c_5\sqrt{nd^2\overline{r}^2 \log n} + \int_{c_5\sqrt{nd^2\overline{r}^2 \log n}}^{\infty} e^{-c_4 t\sqrt{nd^2\overline{r}^2 \log n}} dt \\
&\le c_6 d\overline{r}\sqrt{n \log n}, \qquad (92)
\end{aligned}$$

where c_6 is a positive constant. □

C Proof of Lemma 2

Proof. First, we introduce a directional version of the marginal tail function:

$$H_\xi(E;\boldsymbol{b}) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\{|\langle \boldsymbol{b}_k,\boldsymbol{u}\rangle| \geq \xi\}, \text{ for } \boldsymbol{u}\in E \text{ and } \xi > 0. \tag{93}$$

Lyapunov's inequality and Markov's inequality give the following bounds

$$\left(\frac{1}{QK}\sum_{i=1}^{Q}\sum_{k=1}^{K}|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle|^2\right)^{\frac{1}{2}} \geq \frac{1}{QK}\sum_{i=1}^{Q}\sum_{k=1}^{K}|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \frac{\xi}{QK}\sum_{i=1}^{Q}\sum_{k=1}^{K}\mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi), \tag{94}$$

where we write $\mathbb{1}(A)$ for the $0-1$ random variable that indicates whether the event A takes place. Add and subtract $H_{2\xi}(E;\boldsymbol{b})$ inside the sum, and then take the infimum over $\boldsymbol{u}\in E$ to reach the inequality

$$\inf_{\boldsymbol{u}\in E}\left(\frac{1}{QK}\sum_{i=1}^{Q}\sum_{k=1}^{K}|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle|^2\right)^{\frac{1}{2}} \geq \xi \inf_{\boldsymbol{u}\in E} H_{2\xi}(E;\boldsymbol{b}) - \frac{\xi}{Q}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\left[H_{2\xi}(E;\boldsymbol{b}) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi)\right]. \tag{95}$$

Observe that each summand over index i at the RHS is independent and bounded in magnitude by 1. Therefore, based on [113, Section 6.1], we have

$$\begin{aligned}&\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\left[H_{2\xi}(E;\boldsymbol{b}) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi)\right]\\ &\leq \mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\left[H_{2\xi}(E;\boldsymbol{b}) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi)\right] + t\sqrt{Q},\end{aligned} \tag{96}$$

with probability at least $1-e^{-\frac{t^2}{2}}$.

Next, we simplify the expected supremum. Introduce a soft indicator function:

$$\phi_\xi : \mathbb{R}\to[0,1] \text{ where } \phi_\xi(s) = \begin{cases}0, & |s|\leq\xi,\\ (|s|-\xi)/\xi, & \xi<|s|\leq 2\xi,\\ 1, & 2\xi<|s|.\end{cases}$$

According to [42], we can derive

$$\begin{aligned}&\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\left[H_{2\xi}(E;\boldsymbol{b}) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi)\right]\\ =\ &\frac{1}{K}\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\sum_{k=1}^{K}\left[\mathbb{E}\,\mathbb{1}(|\langle \boldsymbol{b}_k,\boldsymbol{u}\rangle| \geq 2\xi) - \mathbb{1}(|\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle| \geq \xi)\right]\\ \leq\ &\frac{1}{K}\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\left[\mathbb{E}\sum_{k=1}^{K}\phi_\xi\left(\langle \boldsymbol{b}_k,\boldsymbol{u}\rangle\right) - \sum_{k=1}^{K}\phi_\xi\left(\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle\right)\right]\\ \leq\ &\frac{2}{K}\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\epsilon_i\sum_{k=1}^{K}\phi_\xi(\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle)\\ \leq\ &\frac{2}{\xi K}\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\sum_{k=1}^{K}\epsilon_i\langle \boldsymbol{b}_{i,k},\boldsymbol{u}\rangle,\end{aligned} \tag{97}$$

where in the first equation, we write the marginal tail function as an expectation, and then we bound the two indicators using the soft indicator function. In the second inequality, where $\epsilon_i, i=1,\ldots,Q$ are independent Rademacher random variables that are independent from everything else, we use the Giné–Zinn symmetrization [114, Lemma 2.3.1] due to the independence of $\sum_{k=1}^{K} \phi_\xi \left(\langle \boldsymbol{b}_{i,k}, \boldsymbol{u}\rangle\right)$ for $i = 1, \dots, Q$. In the last line, due to the contraction of $\xi\phi_\xi$, we apply the Rademacher comparison principle [115, Eqn.(4.20)].

Hence, we have

$$\inf_{\boldsymbol{u}\in E} \left(\frac{1}{QK}\sum_{i=1}^{Q}\sum_{k=1}^{K} |\langle \boldsymbol{b}_{i,k}, \boldsymbol{u}\rangle|^2\right)^{\frac{1}{2}} \geq \xi \inf_{\boldsymbol{u}\in E} H_{2\xi}(E;\boldsymbol{b}) - \frac{\xi}{Q}\left[\frac{2}{\xi K}\,\mathbb{E}\sup_{\boldsymbol{u}\in E}\sum_{i=1}^{Q}\sum_{k=1}^{K}\epsilon_i\langle \boldsymbol{b}_{i,k}, \boldsymbol{u}\rangle + t\sqrt{Q}\right]. \tag{98}$$

Letting $\boldsymbol{h} = \frac{1}{\sqrt{QK}}\sum_{i=1}^{Q}\sum_{k=1}^{K}\epsilon_i \boldsymbol{b}_{i,k}$, we can finally obtain

$$\inf_{\boldsymbol{u}\in E} \left(\sum_{i=1}^{Q}\sum_{k=1}^{K} |\langle \boldsymbol{b}_{i,k}, \boldsymbol{u}\rangle|^2\right)^{\frac{1}{2}} \geq \xi\sqrt{QK} \inf_{\boldsymbol{u}\in E} H_{2\xi}(E;\boldsymbol{b}) - 2\,\mathbb{E}\sup_{\boldsymbol{u}\in E}\langle \boldsymbol{h}, \boldsymbol{u}\rangle - t\xi\sqrt{K}. \tag{99}$$

This completes the proof of Lemma 2. □

D Proof of Theorem 4

Proof. We prove Theorem 4 using the modified Mendelson's small ball method. Let $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K\}$ be the first K columns of a randomly generated Haar distributed unitary matrix, and let $\{\boldsymbol{\phi}_{i,1}, \dots, \boldsymbol{\phi}_{i,K}\}_{i=1}^{Q}$ be independent copies of $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K\}$. According to Lemma 2, we need to bound

$$H_\xi(\overline{\mathbb{X}}_{\overline{r}}) = \inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\{|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle| \geq \xi\} \tag{100}$$

and

$$W(\overline{\mathbb{X}}_{\overline{r}}) = \mathbb{E}\sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{\sqrt{QK}}\sum_{i=1}^{Q}\sum_{k=1}^{K}\langle \epsilon_i\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle, \tag{101}$$

where $\epsilon_i, i = 1, \dots, Q$ are indepdent Rademacher random variables. Below we study the two quantities separately.

- Lower bound of $H_\xi(\overline{\mathbb{X}}_{\overline{r}})$: As in Appendix B, we also use the Paley-Zygmund inequality (Lemma 5) to bound $H_\xi(\overline{\mathbb{X}}_{\overline{r}})$. Specifically,

$$\begin{aligned} H_\xi(\overline{\mathbb{X}}_{\overline{r}}) &= \inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left(|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle| \geq \xi\right)\\ &= \inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left(|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2 \geq \xi^2\right)\\ &\geq \inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}\left(|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2 \geq \frac{1}{2}\,\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2]\right)\\ &\geq \inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{K}\sum_{k=1}^{K}\frac{(\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2])^2}{4\,\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^4]} \geq c_0,\ \forall\xi \leq \sqrt{\frac{1}{2}\,\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2]}, \end{aligned} \tag{102}$$

  where the first inequality follows because $\mathbb{P}\left(|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2 \geq \xi^2\right)$ is a decreasing function with respect to ξ, the second inequality uses the Paley-Zygmund inequality (Lemma 5) for $|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2$, and the last inequality uses Lemma 6. Below we show that $\left|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle\right|$ is a subexponential random variable and hence satisfies the requirements for both Lemma 5 and Lemma 6. According to the process of Gram-Schmidt orthogonalization for obtaining a Haar-distributed unitary matrix, $\boldsymbol{\phi}_1$ can be obtained by normalizing a standard normal random vector from the distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_{d^N})$. Using $\|\sqrt{d^n}\boldsymbol{\phi}_1\|_{\psi_2} \le O(1)$ [116], we have

$$\|\langle \boldsymbol{\phi}_1\boldsymbol{\phi}_1^H, \boldsymbol{\rho}\rangle\|_{\psi_1} \le \frac{1}{d^n}\|\sqrt{d^n}\boldsymbol{\phi}_1\|_{\psi_2}^2 \|\boldsymbol{\rho}\|_F = O(\frac{1}{d^n}) \tag{103}$$

and hence $\left|\langle \boldsymbol{\phi}_1\boldsymbol{\phi}_1^H, \boldsymbol{\rho}\rangle\right|$ is a subexponential random variable. Finally according to [117], because all the entries in a Haar-distributed unitary matrix have the same distribution due to the translation invariance of Lemma 8, we conclude that each $\left|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle\right|$ is a subexponential random variable for all $k = 1, \dots, d^n$.

To complete the proof of this part, we now study $\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2]$, controlling the upper bound of ξ in (102). Towards that goal, for any $\boldsymbol{\rho} \in \overline{\mathbb{X}}_{\overline{r}}$, we denote its eigenvalue decomposition by $\boldsymbol{\rho} = \sum_{i=1}^{d^n} \lambda_i \boldsymbol{u}_i\boldsymbol{u}_i^H$, where $\{\boldsymbol{u}_i\}$ are unitary vectors and $\{\lambda_i\}$ are the eigenvalues with $\sum_{i=1}^{d^n} \lambda_i^2 = 1$. Now, following (35), we have

$$\begin{aligned}\mathbb{E}[|\langle \boldsymbol{\phi}_k\boldsymbol{\phi}_k^H, \boldsymbol{\rho}\rangle|^2] &= \sum_{j=1}^{d^n}\sum_{l=1}^{d^n} \frac{\lambda_j\lambda_l}{d^{2n}} + \sum_{l=1}^{d^n} \frac{d^n - 1}{d^{2n}(d^n+1)}\lambda_l^2 \\ &= \frac{(\sum_l \lambda_l)^2}{d^{2n}} + \frac{d^n-1}{d^{2n}(d^n+1)}\|\boldsymbol{\rho}\|_F^2 \\ &\ge \frac{d^n-1}{d^{2n}(d^n+1)}. \end{aligned} \tag{104}$$

This together with (102) further implies that

$$H_\xi(\overline{\mathbb{X}}_{\overline{r}}) \ge c_0, \ \forall \xi \le \frac{c_1}{d^n} \tag{105}$$

for some positive constant c_1.

- Upper bound of $W(\overline{\mathbb{X}}_{\overline{r}})$: Since each $\langle \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle$ is a subexponetial random variable with $\|\langle \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle\|_{\psi_1} = O(\frac{1}{d^n})$ according to (103), $\epsilon_i\boldsymbol{\phi}_{i,k}^H\boldsymbol{\rho}\boldsymbol{\phi}_{i,k}$ is a centered subexponential random variable with the subexponential norm $\|\epsilon_i\boldsymbol{\phi}_{i,k}^H\boldsymbol{\rho}\boldsymbol{\phi}_{i,k}\|_{\psi_1} = O(\frac{1}{d^n})$. On the other hand, for any i, the random vectors $\boldsymbol{\phi}_{i,k}$ and $\boldsymbol{\phi}_{i,k'}$ are not dependent to each other for $k \ne k'$. Thus, we use Lemma 11 to obtain its concentration inequality as

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{\sqrt{QK}}\sum_{i=1}^{Q}\sum_{k=1}^{K}\langle \epsilon_i\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle \ge t\right) \\ \le\ & \begin{cases} \left(\frac{4+\epsilon}{\epsilon}\right)^{d^2r_1 + \sum_{i=2}^{n-1} d^2 r_{i-1}r_i + d^2 r_{n-1}} e^{-\frac{c_2 d^{2n}t^2}{4K}}, & t \le \frac{c_4\sqrt{QK}}{d^n} \\ \left(\frac{4+\epsilon}{\epsilon}\right)^{d^2r_1 + \sum_{i=2}^{n-1} d^2 r_{i-1}r_i + d^2 r_{n-1}} e^{-\frac{c_3\sqrt{Q}d^n t}{2\sqrt{K}}}, & t > \frac{c_4\sqrt{QK}}{d^n} \end{cases} \\ \le\ & \begin{cases} e^{-\frac{c_2 d^{2n}t^2}{4K} + Cnd^2\overline{r}^2\log n}, & t \le \frac{c_4\sqrt{QK}}{d^n} \\ e^{-\frac{c_3\sqrt{Q}d^n t}{2\sqrt{K}} + Cnd^2\overline{r}^2\log n}, & t > \frac{c_4\sqrt{QK}}{d^n} \end{cases} \\ \le\ & e^{-\min\{\frac{c_2 d^{2n}t^2}{4K}, \frac{c_3\sqrt{Q}d^n t}{2\sqrt{K}}\} + Cnd^2\overline{r}^2\log n}, \end{aligned} \tag{106}$$

where $\epsilon = \frac{1}{2n}$ is chosen, $\overline{r} = \max_{i=1,\dots,n-1} r_i$, and c_2, c_3, c_4, C are universal constants. Following the same analysis of (92), when $Q = \Omega(nd^2\overline{r}^2 \log n)$, we have

$$W(\overline{\mathbb{X}}_{\overline{r}}) = \mathbb{E}\sup_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \frac{1}{\sqrt{QK}}\sum_{i=1}^{Q}\sum_{k=1}^{K}\langle \epsilon_i\boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle \le c_5\frac{\sqrt{K}d\overline{r}\sqrt{n\log n}}{d^n}, \tag{107}$$

where c_5 is a universal constant.

- Contraction: Combining (105) and (107), and setting $t = \frac{c_0\sqrt{Q}}{2}$, $\xi = \frac{c_1}{d^n}$, and $Q \geq \frac{64c_5^2 n d^2 \overline{r}^2(\log n)}{c_0^2 c_1^2}$, we get

$$\begin{aligned}\inf_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \left(\sum_{i=1}^{Q}\sum_{k=1}^{K} |\langle \boldsymbol{\phi}_{i,k}\boldsymbol{\phi}_{i,k}^H, \boldsymbol{\rho}\rangle|^2\right)^{\frac{1}{2}} &\geq \xi\sqrt{QK}H_\xi(\overline{\mathbb{X}}_{\overline{r}}) - 2W(\overline{\mathbb{X}}_{\overline{r}}) - t\xi\sqrt{K} \\ &\geq \frac{c_0c_1\sqrt{QK}}{d^n} - 2c_5\frac{\sqrt{K}d\overline{r}\sqrt{n\log n}}{d^n} - \frac{c_1\sqrt{QK}}{d^n} \\ &\geq \frac{c_0c_1\sqrt{QK}}{4d^n} \end{aligned} \tag{108}$$

  with probability $1 - e^{-\Omega(Q)}$.

This completes the proof of Theorem 4. □

E Proof of Theorem 5

Proof. Before deriving Theorem 5, we restate our model. We first randomly generate Q Haar distributed unitary matrices $\begin{bmatrix}\boldsymbol{\phi}_{i,1} & \cdots & \boldsymbol{\phi}_{i,d^n}\end{bmatrix}$, which induce Q POVMs of form $\{\boldsymbol{\phi}_{i,1}\boldsymbol{\phi}_{i,1}^H, \dots, \boldsymbol{\phi}_{i,d^n}\boldsymbol{\phi}_{i,d^n}^H\}$, $i = 1, \dots, Q$. Recalling (30) and (31), we have population measurements for the unknown quantum state $\boldsymbol{\rho}^\star$ and total empirical measurements given by $\boldsymbol{p}^Q = \mathcal{A}^Q(\boldsymbol{\rho}^\star)$ and $\widehat{\boldsymbol{p}}^Q$. We then define the statistical measurement error as

$$\boldsymbol{\eta} = \widehat{\boldsymbol{p}}^Q - \boldsymbol{p}^Q = \widehat{\boldsymbol{p}}^Q - \mathcal{A}^Q(\boldsymbol{\rho}^\star) = \left[\boldsymbol{\eta}_1^T, \cdots, \boldsymbol{\eta}_Q^T\right]^T, \tag{109}$$

where $\eta_{i,k}$ is the k-th element in $\boldsymbol{\eta}_i$. With $\widehat{\boldsymbol{p}}^Q$, we estimate the unknown state $\boldsymbol{\rho}^\star$ by solving the following constrained least-squares problem

$$\widehat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}\in\mathbb{X}_{\overline{r}}} \|\mathcal{A}^Q(\boldsymbol{\rho}) - \widehat{\boldsymbol{p}}^Q\|_2^2. \tag{110}$$

Following (38), we have

$$\|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\|_2^2 \leq 2\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle. \tag{111}$$

According to Theorem 4, given $Q \gtrsim nd^2\overline{r}^2(\log n)$, with probability at least $1 - e^{-c_1 Q}$, we have $\|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\|_2^2 \gtrsim \frac{Q}{d^n}\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F^2$. Next, we will upper bound $\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle$. Towards that goal, we first rewrite this term as

$$\langle \boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\rangle = \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H(\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star)\boldsymbol{\phi}_{i,k} \leq \|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\|_F \max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\boldsymbol{\rho}\boldsymbol{\phi}_{i,k}. \tag{112}$$

The rest of the proof is to bound $\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\boldsymbol{\rho}\boldsymbol{\phi}_{i,k}$, which will be achieved by using a covering argument. First, when conditioned on $\{\boldsymbol{\phi}_{i,k}, \forall i, k\}$, we consider any fixed value of $\widetilde{\boldsymbol{\rho}}$ and apply Lemma 3 to establish a concentration inequality for the expression $\sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}$. Denote the event $F := \{\max_{i,k}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}| \lesssim \frac{\log Q + n\log d}{d^n}\}$ which holds with probability $\mathbb{P}(F) = 1 - e^{-c_2(\log Q + n\log d)}$ (its proof is given in Section E.1). Then we have

$$\mathbb{P}\left(\sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \geq t \middle| F\right) \leq 2e^{-\frac{d^{2n}Mt^2}{c_3Q(\log Q + n\log d)^2}}, \tag{113}$$

where c_2 and c_3 are positive constants. The formal proof of (113) is given in Section E.1.

Following the same analysis as in Appendix B.1, there exists an ϵ-net $\widetilde{\mathbb{X}}_{\overline{r}}$ of $\overline{\mathbb{X}}_{\overline{r}}$ such that

$$\begin{aligned}\mathbb{P}\left(\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}} \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\boldsymbol{\rho}\boldsymbol{\phi}_{i,k} \geq t \middle| F\right) &\leq \mathbb{P}\left(\max_{\widetilde{\boldsymbol{\rho}}\in\widetilde{\mathbb{X}}_{\overline{r}}} \sum_{i=1}^{Q}\sum_{k=1}^{d^n} \eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \geq \frac{t}{2} \middle| F\right) \\ &\leq \left(\frac{4+\epsilon}{\epsilon}\right)^{d^2r_1 + \sum_{i=2}^{n-1} d^2r_{i-1}r_i + d^2r_{n-1}} e^{-\frac{d^{2n}Mt^2}{c_3Q(\log Q + n\log d)^2} + \log 2} \\ &\leq e^{-\frac{d^{2n}Mt^2}{c_3Q(\log Q + n\log d)^2} + Cnd^2\overline{r}^2\log n + \log 2}, \end{aligned} \tag{114}$$

where $\epsilon = \frac{1}{2n}$ is chosen, $\overline{r} = \max_{i=1,\dots,n-1} r_i$, and C is a universal constant in the last line. By taking $t = \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n}$ in the above equation, we further obtain

$$\mathbb{P}\left(\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}}\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \le \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n}\middle| F\right) \le 1 - e^{-c_5 nd^2\overline{r}^2\log n}, \tag{115}$$

where c_4 and c_5 are constants.

Now plugging in the probability for the event F, we finally get

$$\begin{aligned}
&\mathbb{P}\left(\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}}\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \le \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n}\right)\\
\ge\ &\mathbb{P}\left(\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}}\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \le \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n} \cap F\right)\\
=\ &\mathbb{P}\left(F\right)\mathbb{P}\left(\max_{\boldsymbol{\rho}\in\overline{\mathbb{X}}_{\overline{r}}}\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \le \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n}\middle| F\right)\\
\ge\ &(1 - e^{-c_2\log(Qd^n)})(1 - e^{-c_5 nd^2\overline{r}^2\log n}) \ge 1 - e^{-c_2(\log Q + n\log d)} - e^{-c_5 nd^2\overline{r}^2\log n}.
\end{aligned} \tag{116}$$

Hence, for $\langle\boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star)\rangle$ in (112), the above equation implies that with probability at least $1 - e^{-c_2(\log Q + n\log d)} - e^{-c_5 nd^2\overline{r}^2\log n}$,

$$\langle\boldsymbol{\eta}, \mathcal{A}^Q(\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star)\rangle \le \frac{c_4\sqrt{Qn\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{M}d^n}\|\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star\|_F. \tag{117}$$

Combining this together with $\|\mathcal{A}^Q(\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star)\|_2^2 \gtrsim \frac{Q}{d^n}\|\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star\|_F^2$, we finally obtain

$$\|\widehat{\boldsymbol{\rho}}-\boldsymbol{\rho}^\star\|_F \lesssim \frac{\sqrt{n\log n}d\overline{r}(\log Q + n\log d)}{\sqrt{MQ}}. \tag{118}$$

This completes the proof of Theorem 5. □

E.1 Proof of (113)

Proof. Conditioning on $\{\boldsymbol{\phi}_{i,k}, \forall i, k\}$, we use Lemma 14 by setting $a_{i,k} = \boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}$ to get

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{Q}\sum_{k=1}^{d^n}\eta_{i,k}\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k} \ge t\middle|\{\boldsymbol{\phi}_{i,k}, \forall i, k\}\right)\\
\le\ & e^{-\frac{Mt}{4\max_{i,k}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|}\min\left\{1, \frac{\max_{i,k}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|t}{4\sum_{i=1}^{Q}\sum_{k=1}^{d^n}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|^2 p_{i,k}}\right\}} + e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{d^n}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|^2 p_{i,k}}}\\
=\ & e^{-\frac{Mt^2}{16\sum_{i=1}^{Q}\sum_{k=1}^{d^n}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|^2 p_{i,k}}} + e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{d^n}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|^2 p_{i,k}}},
\end{aligned} \tag{119}$$

where without loss of generality, we assume that $\frac{\max_{i,k}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|t}{4\sum_{i=1}^{Q}\sum_{k=1}^{d^n}|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}|^2 p_{i,k}} \le 1$ in the last line.

Notice that for any i and k, $\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}$ is a subexponential random variable with subexponential norm $\|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}\|_{\psi_1} = O(\frac{1}{d^n})$ according to (103). By using the concentration equality for the tail of a subexponential random variable [118, Proposition 3], which states that $\mathbb{P}\left(|X| \ge t\right) \le 2e^{-\frac{c_0 t}{\|X\|_{\psi_1}}}$ for any subexponential random variable X with a universal constant c_0, we have

$$\mathbb{P}\left(|\boldsymbol{\phi}_{i,k}^H\widetilde{\boldsymbol{\rho}}\boldsymbol{\phi}_{i,k}| \ge t\right) \le e^{1-c_1 d^n t}, \tag{120}$$

where c_1 is a universal constant. It follows that

$$\mathbb{P}\left(\max_{i,k} |\boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k}| \leq t\right) \geq 1 - Qd^n e^{1-c_1 d^n t} = 1 - e^{1-c_1 d^n t + \log(Qd^n)}. \tag{121}$$

Thus, setting $t = \frac{c_2 \log(Qd^n)}{d^n}$, we obtain

$$\max_{i,k} |\boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k}| \leq \frac{c_2 \log(Qd^n)}{d^n} = \frac{c_2(\log Q + n \log d)}{d^n}, \tag{122}$$

with the probability at least $1 - e^{-c_3 \log(Qd^n)}$, where c_2 and c_3 are positive constants. Now under the event $F = \{\max_{i,k} |\boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k}| \lesssim \frac{\log Q + n \log d}{d^n}\}$, we have $\sum_{i=1}^{Q} \sum_{k=1}^{d^n} |\boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k}|^2 p_{i,k} \leq \max_{i,k} |\boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k}|^2 \sum_{i=1}^{Q} \sum_{k=1}^{d^n} p_{i,k} \leq \frac{c_2^2 Q(\log Q + n \log d)^2}{d^{2n}}$, and thus

$$\mathbb{P}\left(\sum_{i=1}^{Q} \sum_{k=1}^{d^n} \eta_{i,k} \boldsymbol{\phi}_{i,k}^H \widetilde{\boldsymbol{\rho}} \boldsymbol{\phi}_{i,k} \geq t \middle| F\right) \leq 2e^{-\frac{d^{2n} M t^2}{16 c_2^2 Q (\log Q + n \log d)^2}}. \tag{123}$$

□

F Auxiliary Materials

Lemma 5. *([119, Lemma 7.16] Paley-Zygmund inequality) If a nonnegative random variable Z has finite second moment, then we have*

$$\mathbb{P}(Z > t) \geq \frac{(\mathbb{E}\, Z - t)^2}{\mathbb{E}\, Z^2}, 0 \leq t \leq \mathbb{E}\, Z. \tag{124}$$

Lemma 6. *([115, Lemma 3.7]) Let d be an integer and let Z be a positive random variable. Then the following are equivalent:*

- *there is a constant C such that for any $p \geq 2$,*

$$\left(\mathbb{E}[Z^p]\right)^{1/p} \leq C p^{d/2} \left(\mathbb{E}[Z^2]\right)^{1/2}; \tag{125}$$

- *for some $\alpha > 0$,*

$$\mathbb{E} \exp(\alpha Z^{2/d}) < \infty. \tag{126}$$

Lemma 7. *([116, Theorem 2.8.2]) Let $X_1, \ldots, X_N$ be independent, mean zero, subexponential random variables, and $\boldsymbol{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^2 \|\boldsymbol{a}\|_2^2}, \frac{t}{K \|\boldsymbol{a}\|_\infty}\right)\right). \tag{127}$$

where $K = \max_i \|X_i\|_{\psi_1} = \sup_{q \geq 1} \mathbb{E}(|X_i|^q)^{1/q}/q$ and c is a positive constant.

Lemma 8. *([117, Lemma 2.2]) A Haar-distributed random unitary matrix $\boldsymbol{U} \in \mathbb{C}^{D \times D}$ can be equivalently generated by applying the Gram-Schmidt orthogonalization procedure to D independent random vectors $\boldsymbol{z}_i \in \mathbb{C}^D, i = 1, 2, ..., D$), where the entries $z_{i,j}$ are mutually independent standard complex normal random variables.*

Lemma 9. *([120, Corollary 1.2]) Let u_{ij} be an element of $n \times n$ Haar-distributed random unitary matrix $\boldsymbol{U}$. We have*

$$\mathbb{E}[|u_{ij}|^{2d}] = \frac{d!}{n(n+1) \cdots (n+d-1)}. \tag{128}$$

Lemma 10. *For any $\boldsymbol{A}_i, \boldsymbol{A}_i^\star \in \mathbb{R}^{r_{i-1}\times r_i}, i = 1, \dots, N$, we have*

$$\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star = \sum_{i=1}^N \boldsymbol{A}_1^\star\cdots\boldsymbol{A}_{i-1}^\star(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}\cdots\boldsymbol{A}_N. \tag{129}$$

Proof. We expand $\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star$ as

$$\begin{aligned}
& \boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star \\
= \ & \boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2\boldsymbol{A}_3\cdots\boldsymbol{A}_N + \boldsymbol{A}_1^\star\boldsymbol{A}_2\boldsymbol{A}_3\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star \\
= \ & (\boldsymbol{A}_1 - \boldsymbol{A}_1^\star)\boldsymbol{A}_2\cdots\boldsymbol{A}_N + \boldsymbol{A}_1^\star\boldsymbol{A}_2\boldsymbol{A}_3\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\boldsymbol{A}_3\cdots\boldsymbol{A}_N + \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\boldsymbol{A}_3\cdots\boldsymbol{A}_N - \boldsymbol{A}_1^\star\boldsymbol{A}_2^\star\cdots\boldsymbol{A}_N^\star \\
= \ & \cdots = \sum_{i=1}^N \boldsymbol{A}_1^\star\cdots\boldsymbol{A}_{i-1}^\star(\boldsymbol{A}_i - \boldsymbol{A}_i^\star)\boldsymbol{A}_{i+1}\cdots\boldsymbol{A}_N.
\end{aligned} \tag{130}$$

□

Lemma 11. *[121, Theorem 3.1] Suppose that $X = \sum_{i=1}^Q\sum_{k=1}^K w_k X_{i,k}$, where $w_k, k = 1, \dots, K$ are constants, and each $X_{i,k}, i = 1, \dots, Q, k = 1, \dots, K$ is a zero-mean, subexponential random variable with $\|X_{i,k}\|_{\psi_1}$. In addition, the Q multivariate random variables $(X_{i,1}, \dots, X_{i,K}), i = 1, \dots, Q$ are mutually independent. However, it is possible for the variables $X_{i,k}$ and $X_{i,k'}, k' \neq k$ within each multivariate random variable to be dependent. Then*

$$\mathbb{P}\left(X > t\right) \leq \begin{cases} e^{-\frac{t^2}{4T^2}}, & t \leq 2T^2H, \\ e^{-\frac{tH}{2}}, & t > 2T^2H. \end{cases} \tag{131}$$

where $T = \sum_{k=1}^K w_k\sqrt{\sum_{i=1}^Q c_{i,k}\|X_{i,k}\|_{\psi_1}^2}$ and $H = \left(\min_k \frac{\sqrt{\sum_{i=1}^Q c_{i,k}\|X_{i,k}\|_{\psi_1}^2}}{\sum_{k=1}^K w_k\sqrt{\sum_{i=1}^Q c_{i,k}\|X_{i,k}\|_{\psi_1}^2}}\right)\cdot\left(\min_i\{\frac{d_{i,k}}{\|X_{i,k}\|_{\psi_1}}\}\right)$ with constants $c_{i,k}$ and $d_{i,k}$.

Below, we extend the concentration bounds presented in [103, Lemmas 2&3] for a single multinomial random variable to encompass multiple multinomial random variables.

Lemma 12. *Suppose that the Q multivariate random variables $(f_{i,k}, \dots, f_{i,K}), i = 1, \dots, Q$ are mutually independent and follow the multinomial distribution* $\text{Multinomial}(M, \boldsymbol{p}_i)$ *with $\sum_{k=1}^K f_{i,k} = M$ and $\boldsymbol{p}_i = [p_{i,1}\dots, p_{i,K}]$, respectively. Let $a_{i,1}, \dots, a_{i,K} \geq 0$ be fixed such that $\sum_{k=1}^K a_{i,k}p_{i,k} \neq 0, i = 1, \dots, Q$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}\big(\frac{f_{i,k}}{M} - p_{i,k}\big) > t\right) \leq e^{-\frac{Mt}{2a_{\max}}\min\left\{1, \frac{a_{\max}t}{2\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}^2 p_{i,k}}\right\}}, \tag{132}$$

where $a_{\max} = \max_{i,k} a_{i,k}$.

Proof. For any $v > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}\big(\frac{f_{i,k}}{M} - p_{i,k}\big) > t\right) &= \mathbb{P}\left(v\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}\frac{f_{i,k}}{M} > v\big(t + \sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k}\big)\right) \\
&\leq \mathbb{P}\left(e^{v\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}\frac{f_{i,k}}{M}} \geq e^{v(t+\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k})}\right) \\
&\leq e^{-v(t+\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k})}\,\mathbb{E}\,e^{v\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}\frac{f_{i,k}}{M}} \\
&= e^{-v(t+\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k})}\prod_{i=1}^Q\mathbb{E}\,e^{v\sum_{k=1}^K a_{i,k}\frac{f_{i,k}}{M}} \\
&\leq e^{-v(t+\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k})}e^{v\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}p_{i,k}+\sum_{i=1}^Q\sum_{k=1}^K p_{i,k}\frac{v^2a_{i,k}^2}{M}} \\
&= e^{-vt+\sum_{i=1}^Q\sum_{k=1}^K p_{i,k}\frac{v^2a_{i,k}^2}{M}} \\
&\leq e^{-\frac{Mt}{2a_{\max}}\min\left\{1, \frac{a_{\max}t}{2\sum_{i=1}^Q\sum_{k=1}^K a_{i,k}^2 p_{i,k}}\right\}},
\end{aligned} \tag{133}$$

where the second inequality uses Markov's inequality, the fourth line follows from the independence of multivariate random variables $(f_{i,k}, \ldots, f_{i,K}), i = 1, \ldots, Q$, the third inequality utilizes [103, Lemma 2] for $\mathbb{E}\, e^{v \sum_{k=1}^{K} a_{i,k} \frac{f_{i,k}}{M}}$, and the last line follows by setting $v = \frac{Mt}{2\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}} \leq \frac{M}{a_{\max}}$ when $t \leq \frac{2\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}{a_{\max}}$ and $v = \frac{M}{a_{\max}}$ when $t \geq \frac{2\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}{a_{\max}}$. □

Lemma 13. *Suppose that the Q multivariate random variables $(f_{i,k}, \ldots, f_{i,K}), i = 1, \ldots, Q$ are mutually independent and follow the multinomial distribution* $\mathrm{Multinomial}(M, \boldsymbol{p}_i)$ *with $\sum_{k=1}^{K} f_{i,k} = M$ and $\boldsymbol{p}_i = [p_{i,1} \cdot \ldots, p_{i,K}]$, respectively. Let $a_{i,1}, \ldots, a_{i,K} \geq 0$ be fixed such that $\sum_{k=1}^{K} a_{i,k} p_{i,k} \neq 0, i = 1, \ldots, Q$. Then, for any $t > 0$,*

$$\mathbb{P}\left(-\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > t\right) \leq e^{-\frac{Mt^2}{2\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}}. \tag{134}$$

Proof. Following the same approach for proving Lemma 12, for any $v < 0$, we have

$$\begin{aligned} \mathbb{P}\left(-\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > t\right) &= \mathbb{P}\left(v\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}\frac{f_{i,k}}{M} > v(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}p_{i,k} - t)\right) \\ &\leq \mathbb{P}\left(e^{v\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}\frac{f_{i,k}}{M}} \geq e^{v(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}p_{i,k} - t)}\right) \\ &\leq e^{-v(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}p_{i,k} - t)}\, \mathbb{E}\, e^{v\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}\frac{f_{i,k}}{M}} \\ &= e^{-v(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}p_{i,k} - t)} \Pi_{i=1}^{Q}\, \mathbb{E}\, e^{v\sum_{k=1}^{K} a_{i,k}\frac{f_{i,k}}{M}} \\ &\leq e^{vt + \sum_{i=1}^{Q}\sum_{k=1}^{K} p_{i,k}\frac{a_{i,k}^2 v^2}{2M}} \\ &= e^{-\frac{Mt^2}{2\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}}, \end{aligned} \tag{135}$$

where the derivations before the last line are the same as those for proving Lemma 12 and the last line follows by setting $v = -\frac{tM}{\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}$. □

Lemma 12 and Lemma 13, together leads to the following multinomial concentration bounds.

Lemma 14. *Suppose that the Q multivariate random variables $(f_{i,k}, \ldots, f_{i,K}), i = 1, \ldots, Q$ are mutually independent and follow the multinomial distribution* $\mathrm{Multinomial}(M, \boldsymbol{p}_i)$ *with $\sum_{k=1}^{K} f_{i,k} = M$ and $\boldsymbol{p}_i = [p_{i,1} \cdot \ldots, p_{i,K}]$, respectively. Let $a_{i,1}, \ldots, a_{i,K}$ be fixed. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > t\right) \leq e^{-\frac{Mt}{4a_{\max}}\min\left\{1, \frac{a_{\max}t}{4\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}\right\}} + e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}}}, \tag{136}$$

where $a_{\max} = \max_{i,k} |a_{i,k}|$.

Proof. Since $\{a_{i,k}\}, i = 1, \ldots, Q, k = 1, \ldots, K$ could be positive or negative, we separate the set into three sets P, N and Z such that $a_{i,k} > 0$ for $\{i,k\} \in P$, $a_{i,k} < 0$ for $\{i,k\} \in N$, and $a_{i,k} = 0$ for $\{i,k\} \in Z$. In addition, when $p_{i,k} = 0$, we have $f_{i,k} = 0$ and further obtain $a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) = 0$. Thus, without loss of generality, we assume that

$p_{i,k} > 0$ for all i, k. Now we have

$$\begin{aligned}
& \mathbb{P}\left(\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > t\right) \\
\leq \;& \mathbb{P}\left(\sum_{\{i,k\}\in P} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2} \cup \sum_{\{i,k\}\in N} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\right) \\
\leq \;& \mathbb{P}\left(\sum_{\{i,k\}\in P} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\right) + \mathbb{P}\left(\sum_{\{i,k\}\in N} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\right) \\
= \;& \mathbb{P}\left(\sum_{\{i,k\}\in P} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) + \sum_{\{i,k\}\in N\cup Z} 0\cdot(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\right) \\
& + \mathbb{P}\left(\sum_{\{i,k\}\in N} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) + \sum_{\{i,k\}\in P\cup Z} 0\cdot(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\right) \\
\leq \;& e^{-\frac{Mt}{4\tilde{a}_{\max}}\min\left\{1, \frac{\tilde{a}_{\max}t}{4\sum_{i=1}^{Q}\sum_{k=1}^{K}\tilde{a}_{i,k}^2 p_{i,k}}\right\}} + e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{K}\hat{a}_{i,k}^2 p_{i,k}}} \\
\leq \;& e^{-\frac{Mt}{4a_{\max}}\min\left\{1, \frac{a_{\max}t}{4\sum_{i=1}^{Q}\sum_{k=1}^{K}a_{i,k}^2 p_{i,k}}\right\}} + e^{-\frac{Mt^2}{8\sum_{i=1}^{Q}\sum_{k=1}^{K}a_{i,k}^2 p_{i,k}}}, \qquad (137)
\end{aligned}$$

where the first inequality follows the fact $\{\sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > t\} \subseteq \{\sum_{\{i,k\}\in P} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\} \cup \{\sum_{\{i,k\}\in N} a_{i,k}(\frac{f_{i,k}}{M} - p_{i,k}) > \frac{t}{2}\}$. In addition, we define two sets $\tilde{a}_{i,k} = \begin{cases} a_{i,k}, & \{i,k\}\in P \\ 0, & \{i,k\}\in N\cup Z \end{cases}$ and $\hat{a}_{i,k} = \begin{cases} a_{i,k}, & \{i,k\}\in N \\ 0, & \{i,k\}\in P\cup Z \end{cases}$ in the second inequality, and in the last line we uses two facts that $\tilde{a}_{\max} = \max_{i,k}|\tilde{a}_{i,k}| \leq a_{\max}$ and $\max\left\{\sum_{i=1}^{Q}\sum_{k=1}^{K}\tilde{a}_{i,k}^2 p_{i,k}, \sum_{i=1}^{Q}\sum_{k=1}^{K}\hat{a}_{i,k}^2 p_{i,k}\right\} \leq \sum_{i=1}^{Q}\sum_{k=1}^{K} a_{i,k}^2 p_{i,k}$. □

References

[1] John Preskill. Quantum computing in the nisq era and beyond. *Quantum*, 2:79, 2018.

[2] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. Quantum supremacy using a programmable superconducting processor. *Nature*, 574(7779):505–510, 2019.

[3] Jerry Chow, Oliver Dial, and Jay Gambetta. Ibm quantum breaks the 100-qubit processor barrier. *IBM Research Blog*, 2021.

[4] K Vogel and H Risken. Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. *Physical Review A*, 40(5):2847, 1989.

[5] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[6] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[7] Zdenek Hradil. Quantum-state estimation. *Physical Review A*, 55(3):R1561, 1997.

[8] J Řeháček, Z Hradil, and M Ježek. Iterative algorithm for reconstruction of entangled states. *Physical Review A*, 63(4):040303, 2001.

[9] Robin Blume-Kohout. Optimal, reliable estimation of quantum states. *New Journal of Physics*, 12(4):043034, 2010.

[10] Christopher Granade, Joshua Combes, and DG Cory. Practical bayesian tomography. *new Journal of Physics*, 18(3):033024, 2016.

[11] Joseph M Lukens, Kody JH Law, Ajay Jasra, and Pavel Lougovski. A practical and efficient approach for bayesian quantum state estimation. *New Journal of Physics*, 22(6):063038, 2020.

[12] Robin Blume-Kohout. Robust error bars for quantum tomography. *arXiv:1202.5270*, 2012.

[13] Philippe Faist and Renato Renner. Practical and reliable error bars in quantum tomography. *Physical review letters*, 117(1):010404, 2016.

[14] Anastasios Kyrillidis, Amir Kalev, Dohyung Park, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable compressed sensing quantum state tomography via non-convex methods. *npj Quantum Information*, 4(1):1–7, 2018.

[15] Fernando GSL Brandão, Richard Kueng, and Daniel Stilck França. Fast and robust quantum state tomography from few basis measurements. *arXiv:2009.08216*, 2020.

[16] Giacomo Torlai, Guglielmo Mazzola, Juan Carrasquilla, Matthias Troyer, Roger Melko, and Giuseppe Carleo. Neural-network quantum state tomography. *Nature Physics*, 14(5):447–450, 2018.

[17] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[18] Sanjaya Lohani, Brian T Kirby, Michael Brodsky, Onur Danaci, and Ryan T Glasser. Machine learning assisted quantum state estimation. *Machine Learning: Science and Technology*, 1(3):035007, 2020.

[19] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[20] Madalin Guţă, Jonas Kahn, Richard Kueng, and Joel A Tropp. Fast state tomography with optimal error bounds. *Journal of Physics A: Mathematical and Theoretical*, 53(20):204001, 2020.

[21] Daniel Stilck França, Fernando GS Brandão, and Richard Kueng. Fast and robust quantum state tomography from few basis measurements. In *16th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[22] Vladislav Voroninski. Quantum tomography from few full-rank observables. *arXiv:1309.7669*, 2013.

[23] J Haah, AW Harrow, Z Ji, X Wu, and N Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017.

[24] Yi-Kai Liu. Universal low-rank matrix recovery from pauli measurements. *Advances in Neural Information Processing Systems*, 24, 2011.

[25] J. Eisert, M. Cramer, and M. B. Plenio. Colloquium: Area laws for the entanglement entropy. *Rev. Mod. Phys.*, 82:277–306, Feb 2010.

[26] Kyungjoo Noh, Liang Jiang, and Bill Fefferman. Efficient classical simulation of noisy random quantum circuits in one dimension. *Quantum*, 4:318, 2020.

[27] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[28] Tillmann Baumgratz, David Gross, Marcus Cramer, and Martin B Plenio. Scalable reconstruction of density matrices. *Physical review letters*, 111(2):020401, 2013.

[29] Alexander Lidiak, Casey Jameson, Zhen Qin, Gongguo Tang, Michael B Wakin, Zhihui Zhu, and Zhexuan Gong. Quantum state tomography with tensor train cross approximation. *arXiv:2207.06397*, 2022.

[30] Marcus Cramer, Martin B Plenio, Steven T Flammia, Rolando Somma, David Gross, Stephen D Bartlett, Olivier Landon-Cardinal, David Poulin, and Yi-Kai Liu. Efficient quantum state tomography. *Nature communications*, 1(1):1–7, 2010.

[31] BP Lanyon, C Maier, Milan Holzäpfel, Tillmann Baumgratz, C Hempel, P Jurcevic, Ish Dhand, AS Buyskikh, AJ Daley, Marcus Cramer, et al. Efficient tomography of a quantum many-body system. *Nature Physics*, 13(12):1158–1162, 2017.

[32] Jun Wang, Zhao-Yu Han, Song-Bo Wang, Zeyang Li, Liang-Zhu Mu, Heng Fan, and Lei Wang. Scalable quantum tomography with fidelity estimation. *Physical Review A*, 101(3):032321, 2020.

[33] Frank Verstraete, Juan J Garcia-Ripoll, and Juan Ignacio Cirac. Matrix product density operators: Simulation of finite-temperature and dissipative systems. *Physical review letters*, 93(20):207204, 2004.

[34] Bogdan Pirvu, Valentin Murg, J Ignacio Cirac, and Frank Verstraete. Matrix product operator representations. *New Journal of Physics*, 12(2):025012, 2010.

[35] Albert H Werner, Daniel Jaschke, Pietro Silvi, Martin Kliesch, Tommaso Calarco, Jens Eisert, and Simone Montangero. Positive tensor network approach for simulating open quantum many-body systems. *Physical review letters*, 116(23):237201, 2016.

[36] Jiří Guth Jarkovskỳ, András Molnár, Norbert Schuch, and J Ignacio Cirac. Efficient description of many-body systems with matrix product density operators. *PRX Quantum*, 1(1):010304, 2020.

[37] Fernando GSL Brandao, Aram W Harrow, and Michał Horodecki. Local random quantum circuits are approximate polynomial-designs. *Communications in Mathematical Physics*, 346:397–434, 2016.

[38] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[39] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[40] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[41] Armin Eftekhari and Michael B Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.

[42] Joel A Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance*, pages 67–101. Springer, 2015.

[43] Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

[44] Min Yu, Dongxiao Li, Jingcheng Wang, Yaoming Chu, Pengcheng Yang, Musang Gong, Nathan Goldman, and Jianming Cai. Experimental estimation of the quantum fisher information from randomized measurements. *Physical Review Research*, 3(4):043122, 2021.

[45] Andreas Elben, Steven T Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, 5(1):9–24, 2023.

[46] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings 26*, pages 3–18. Springer, 2015.

[47] Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

[48] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[49] Holger Rauhut, Reinhold Schneider, and Željka Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

[50] Jian-Feng Cai, Jingyang Li, and Dong Xia. Provable tensor-train format tensor completion by riemannian optimization. *Journal of Machine Learning Research*, 23(123):1–77, 2022.

[51] Ivan Oseledets and Eugene Tyrtyshnikov. Tt-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432(1):70–88, 2010.

[52] Dmitry V Savostyanov. Quasioptimality of maximum-volume cross interpolation of tensors. *Linear Algebra and its Applications*, 458:217–244, 2014.

[53] AI Osinsky. Tensor trains approximation estimates in the chebyshev norm. *Computational Mathematics and Mathematical Physics*, 59(2):201–206, 2019.

[54] Sergei A Goreinov and Eugene E Tyrtyshnikov. The maximal-volume concept in approximation by low-rank matrices. *Contemporary Mathematics*, 280:47–52, 2001.

[55] Keaton Hamm and Longxiu Huang. Perspectives on cur decompositions. *Applied and Computational Harmonic Analysis*, 48(3):1088–1099, 2020.

[56] HanQin Cai, Keaton Hamm, Longxiu Huang, and Deanna Needell. Robust cur decomposition: Theory and imaging applications. *SIAM Journal on Imaging Sciences*, 14(4):1472–1503, 2021.

[57] Zhen Qin, Alexander Lidiak, Zhexuan Gong, Gongguo Tang, Michael B Wakin, and Zhihui Zhu. Error analysis of tensor-train cross approximation. In *Thirty-Sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[58] Johann A Bengua, Ho N Phien, Hoang Duong Tuan, and Minh N Do. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Transactions on Image Processing*, 26(5):2466–2479, 2017.

[59] Masaaki Imaizumi, Takanori Maehara, and Kohei Hayashi. On tensor train rank minimization: Statistical efficiency and scalable algorithm. *advances in neural information processing systems*, 30, 2017.

[60] Wenqi Wang, Vaneet Aggarwal, and Shuchin Aeron. Tensor completion by alternating minimization under the tensor train (tt) model. *arXiv:1609.05587*, 2016.

[61] Holger Rauhut, Reinhold Schneider, and Željka Stojanac. Tensor completion in hierarchical tensor representations. In *Compressed sensing and its applications*, pages 419–450. Springer, 2015.

[62] Stanislav Budzinskiy and Nikolai Zamarashkin. Tensor train completion: local recovery guarantees via riemannian optimization. *arXiv:2110.03975*, 2021.

[63] Junli Wang, Guangshe Zhao, Dingheng Wang, and Guoqi Li. Tensor completion using low-rank tensor train decomposition by riemannian optimization. In *2019 Chinese Automation Congress (CAC)*, pages 3380–3384. IEEE, 2019.

[64] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[65] Thomas A Severini. *Elements of distribution theory*, volume 17. Cambridge University Press, 2005.

[66] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[67] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

[68] E. Knill. Approximation by quantum circuits. *arXiv:quant-ph/9508006*, 1995.

[69] Qibin Zhao, Guoxu Zhou, Shengli Xie, Liqing Zhang, and Andrzej Cichocki. Tensor ring decomposition. *arXiv:1606.05535*, 2016.

[70] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. On manifolds of tensors of fixed tt-rank. *Numerische Mathematik*, 120(4):701–731, 2012.

[71] Holger Rauhut, Reinhold Schneider, and Željka Stojanac. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications*, 523:220–262, 2017.

[72] Longhao Yuan, Qibin Zhao, Lihua Gui, and Jianting Cao. High-order tensor completion via gradient-based optimization under tensor train format. *Signal Processing: Image Communication*, 73:53–61, 2019.

[73] Longhao Yuan, Chao Li, Danilo Mandic, Jianting Cao, and Qibin Zhao. Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9151–9158, 2019.

[74] David Perez-García, Frank Verstraete, Michael M Wolf, and J Ignacio Cirac. Matrix product state representations. *Quantum Information and Computation*, 7(5-6):401–430, 2007.

[75] Frank Verstraete and J Ignacio Cirac. Matrix product states represent ground states faithfully. *Physical review b*, 73(9):094423, 2006.

[76] Frank Verstraete, Valentin Murg, and J Ignacio Cirac. Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems. *Advances in physics*, 57(2):143–224, 2008.

[77] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of physics*, 326(1):96–192, 2011.

[78] Matthias Ohliger, Vincent Nesme, and Jens Eisert. Efficient and feasible state tomography of quantum many-body systems. *New Journal of Physics*, 15(1):015024, 2013.

[79] Jose I Latorre. Image compression and entanglement. *quant-ph/0510031*, 2005.

[80] Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. *arXiv:1711.00811*, 2017.

[81] Edwin Stoudenmire and David J Schwab. Supervised learning with tensor networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[82] Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. *Advances in neural information processing systems*, 28, 2015.

[83] Yinchong Yang, Denis Krompass, and Volker Tresp. Tensor-train recurrent neural networks for video classification. In *International Conference on Machine Learning*, pages 3891–3900. PMLR, 2017.

[84] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Compressing recurrent neural network with tensor train. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 4451–4458. IEEE, 2017.

[85] Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. Long-term forecasting using tensor-train rnns. *Arxiv*, 2017.

[86] Xindian Ma, Peng Zhang, Shuai Zhang, Nan Duan, Yuexian Hou, Ming Zhou, and Dawei Song. A tensorized transformer for language modeling. *Advances in neural information processing systems*, 32, 2019.

[87] Evgeny Frolov and Ivan Oseledets. Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(3):e1201, 2017.

[88] Georgii S Novikov, Maxim E Panov, and Ivan V Oseledets. Tensor-train density estimation. In *Uncertainty in Artificial Intelligence*, pages 1321–1331. PMLR, 2021.

[89] Maxim A Kuznetsov and Ivan V Oseledets. Tensor train spectral method for learning of hidden markov models (hmm). *Computational Methods in Applied Mathematics*, 19(1):93–99, 2019.

[90] Alexander Novikov, Pavel Izmailov, Valentin Khrulkov, Michael Figurnov, and Ivan V Oseledets. Tensor train decomposition on tensorflow (t3f). *J. Mach. Learn. Res.*, 21(30):1–7, 2020.

[91] Richard G Baraniuk, Mark A Davenport, Ronald A DeVore, and Michael B Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2007.

[92] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

[93] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[94] Sjoerd Dirksen. Tail bounds via generic chaining. *Electron. J. Probab*, 20(53):1–29, 2015.

[95] Holger Rauhut and Ulrich Terstiege. Low-rank matrix recovery via rank one tight frame measurements. *Journal of Fourier Analysis and Applications*, 25:588–593, 2019.

[96] Joel A Tropp. A comparison principle for functions of a uniformly random subspace. *Probability Theory and Related Fields*, 153:759–769, 2012.

[97] Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):1497–1529, 2006.

[98] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[99] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12:805–849, 2012.

[100] Andrew V Carter. Deficiency distance between multinomial and multivariate normal experiments. *The Annals of Statistics*, 30(3):708–730, 2002.

[101] Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 2009.

[102] Frédéric Ouimet. A precise local limit theorem for the multinomial distribution and some applications. *Journal of Statistical Planning and Inference*, 215:218–233, 2021.

[103] Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness implies generalization via data-dependent generalization bounds. In *International Conference on Machine Learning*, pages 10866–10894. PMLR, 2022.

[104] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

[105] Hengkang Wang, Taihui Li, Zhong Zhuang, Tiancong Chen, Hengyue Liang, and Ju Sun. Early stopping for deep image prior. *arXiv:2112.06074*, 2021.

[106] Lijun Ding, Zhen Qin, Liwei Jiang, Jinxin Zhou, and Zhihui Zhu. A validation approach to over-parameterized matrix and image recovery. *arXiv:2209.10675*, 2022.

[107] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[108] Andrew J Scott. Tight informationally complete quantum measurements. *Journal of Physics A: Mathematical and General*, 39(43):13507, 2006.

[109] Cécilia Lancien and Andreas Winter. Distinguishing multi-partite states by local measurements. *Communications in Mathematical Physics*, 323:555–573, 2013.

[110] Hermine Biermé and Céline Lacaux. Modulus of continuity of some conditionally sub-gaussian fields, application to stable random fields. *Bernoulli*, 21(3):1719–1759, 2015.

[111] Krzysztof Zajkowski. Bounds on tail probabilities for quadratic forms in dependent sub-gaussian random variables. *Statistics & Probability Letters*, 167:108898, 2020.

[112] Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

[113] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[114] Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

[115] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

[116] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[117] Dénes Petz and Júlia Réffy. On asymptotics of large haar distributed unitary matrices. *Periodica Mathematica Hungarica*, 49(1):103–117, 2004.

[118] Junren Chen and Michael K Ng. Error bound of empirical l_2 risk minimization for noisy standard and generalized phase retrieval problems. *arXiv:2205.13827*, 2022.

[119] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing, Applied and Numerical Harmonic Analysis*. Birkhäuser, 2013.

[120] Jonathan Novak. Truncations of random unitary matrices and young tableaux. *math/0608108*, 2006.

[121] Yuta Tanoue. Concentration inequality of sums of dependent subexponential random variables and application to bounds for value-at-risk. *Communications in Statistics-Theory and Methods*, pages 1–20, 2022.