A BLOCK TENSOR TRAIN BURER-MONTEIRO FRAMEWORK FOR LOW-RANK QUANTUM STATE TOMOGRAPHY*

SHAKIR SHOWKAT SOFI†, CHARLOTTE VERMEYLEN†, FATEMEH MOHAMMADI‡, AND LIEVEN DE LATHAUWER†

Abstract. Quantum state tomography is a fundamental technique for estimating the state of a quantum system from measured data and plays a crucial role in evaluating the performance of quantum devices. However, standard estimation methods become computationally prohibitive as the system size increases due to the exponential growth of the density matrix, describing a quantum state, with the number of qubits. We propose a low-rank tensor-network framework for mixed-state quantum state tomography based on a block tensor train (Block-TT) factorization. Specifically, the density matrix is represented as the contraction of a Block-TT with its Hermitian transpose, yielding a TT analogue of the Burer-Monteiro factorization. This parameterization guarantees Hermiticity and positive semidefiniteness by construction while compressing the number of optimization variables from exponential to linear in the number of qubits. Building on this representation, we develop single-site and two-site density matrix renormalization group (DMRG) algorithms for estimating quantum states from compressed measurements. The resulting methods operate directly on the compressed parameterization, support adaptive rank refinement, and exploit efficient tensor-network contractions for expectation-value evaluation. The framework is applicable to a broad class of low-rank quantum states, including pure states, nearly pure states, and ground states that admit accurate tensor-network approximations. Numerical experiments demonstrate accurate state reconstruction from limited measurements together with substantial reductions in memory requirements and computational cost compared with conventional low-rank tomography methods.

Key words. tensor networks, low-rank approximation, tensor trains, compressed sensing, quantum state tomography

AMS subject classifications. 81P18, 15A69, 65F55

1. Introduction.

Quantum computing has attracted significant interest in recent years, driven by emerging tools for analyzing large-scale systems and optimizing over high-dimensional spaces. The study of quantum systems is increasingly important in both academia and industry, with applications spanning quantum communication, sensing, and control. A quantum system's state is fully described by its density matrix—a Hermitian, positive semidefinite (PSD) operator with unit trace. It provides a unified representation of both pure states and mixed states (probabilistic mixtures of pure states), where diagonal entries correspond to state populations and off-diagonal entries encode quantum coherences [44, 36].

A typical quantum algorithm begins by initializing a system in a specific quantum state (state preparation), followed by manipulating the state via quantum gates (state evolution), and finally extracting classical information through measurement (state readout/tomography). Each stage presents its own challenges and, in particular, the

*This paper is an extended version of our conference paper *Tensor Train Quantum State Tomography using Compressed Sensing*, in the 33rd European Signal Processing Conference (EUSIPCO), pp. 1332–1336, 2025. See the contribution statement for details on new additions.

Funding: This work was supported by the Flemish Government's AI Research Program and KU Leuven Internal Funds (iBOF/23/064, C14/22/096). Shakir Showkat Sofi (shakir-showkat.sofi@kuleuven.be), Charlotte Vermeylen (charlotte.vermeylen@kuleuven.be), and Lieven De Lathauwer (lieven.delathauwer@kuleuven.be) are affiliated with Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

† Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium; Group Science, Engineering and Technology, KU Leuven Kulak, Kortrijk, Belgium.

‡ Departments of Computer Science and Mathematics, KU Leuven, Leuven, Belgium. (fatemeh.mohammadi@kuleuven.be).

final step—quantum state tomography (QST)—is fundamentally challenging. QST is an inverse problem: from the measurements, the state cannot be accessed directly; instead, the measurements yield samples from a probability distribution determined by the quantum state. The goal of QST is to reconstruct the quantum state from these samples. Two key challenges arise in this reconstruction problem. First, the number of measurements—and the associated computational cost—required to reconstruct a generic, unstructured state grows exponentially with the number of qubits defining it, an obstacle known as the "curse of dimensionality." Second, the reconstructed state must satisfy the physical constraints of quantum mechanics, namely Hermiticity, positive semidefiniteness, and unit trace, while remaining amenable to efficient storage and post-processing, such as the computation of expectation values. This naturally raises a fundamental question: which quantum states admit efficient representations, and what structural properties enable them? In practice, many quantum states of interest exhibit special structural properties. These include pure or low-entropy states, ground states of local Hamiltonians, product states, and other structured states in which only a subset of entries contains significant information [11, 16, 31]. Many of these examples can be unified under the broader framework of data-sparse or low-rank (in the matrix or tensor sense) structures. QST for such structured states can be performed efficiently by leveraging this structure, as only a partial set of measurements may be sufficient for reconstruction [16, 31].

1.1. Related Work. When the density matrix is known to be (approximately) low rank, *low-rank* QST exploits this structure to reconstruct the state from significantly fewer measurements. Rank constraints can be incorporated either implicitly, by promoting low rank via a rank-minimization objective implemented through the nuclear norm as a convex surrogate under measurement constraints [16, 32, 31, 62], or explicitly, by optimizing a fixed-rank factorized model, as in Burer–Monteiro (Cholesky-like) formulations [4, 27]. Nevertheless, the dependence of the model parameters remains exponential in the number of qubits.

Alternatively, when the density matrix is represented as a high-order tensor, tensor networks—interconnected networks of low-order tensors—are employed to achieve compact representations and alleviate the curse of dimensionality. Prominent tensor network formats include matrix product states (MPS, also known as tensor trains (TT)), tree tensor networks, and projected entangled pair states [46, 42, 39, 22]. In particular, MPS comes with highly efficient optimization schemes based on the density matrix renormalization group (DMRG) [63, 52, 64, 23, 41, 18, 51, 25, 15], making *MPS-based* QST scalable for large-scale quantum many-body systems whose states admit low-rank tensor network representations, including ground states, GHZ states, cluster states, and AKLT states [7, 28, 26, 54]. In [48], QST is generalized to matrix product operators (MPO), extending MPS methods from pure states to noisy mixed states [61]. A key challenge in MPO-based approaches is ensuring that the reconstructed density matrix remains physically valid [61, 8]. Although specialized variants with constraints on MPO cores can enforce positivity [8], designing DMRG-like algorithms in this setting is generally challenging, and such algorithms remain underexplored for the QST task. While there are a few works in this direction, they do not address the original QST problem; instead, they focus on estimating the underlying density matrix from reduced density matrices [7, 17].

A complementary line of work avoids full state reconstruction and instead targets specific structural or functional aspects of the density matrix. For instance, *permutationally invariant* QST efficiently reconstructs the permutationally invariant

component of a state, with measurement complexity scaling only quadratically in the number of qubits, and is particularly well suited for states close to permutation symmetry, such as Dicke and spin-squeezed states [59, 34]. In a different direction, *shadow tomography* constructs compact classical representations (classical shadows) via randomized measurements, enabling efficient estimation of many observables with sample complexity that scales only logarithmically in the number of observables [1, 20]. Extending this idea of partial information extraction, *selective* QST focuses on estimating only a chosen subset of density matrix entries, avoiding full reconstruction [2, 5, 14, 35, 55]. One of the questions in this class of methods is whether such partial information is sufficient to infer reliable information about the full state, or, in other words, whether it is possible to use such methods to guarantee the recoverability of the full state, which is of actual interest in many applications. An algebraic completion-based approach with deterministic recovery conditions leveraging selective measurements has been developed in [55]

1.2. Contributions. In this work, we present a block tensor train (Block-TT) Burer-Monteiro framework for learning high-dimensional mixed quantum states via compressed sensing. By compressing the parameter space from exponential to linear in the number of qubits (or qudits, more generally), the Block-TT Burer-Monteiro representation directly mitigates the curse of dimensionality while inherently ensuring physical state validity by construction. Extending our preliminary conference work [54], where this representation was first introduced, this paper makes the following contributions:

- We develop novel single-site and two-site DMRG algorithms for compressed-sensing quantum state tomography in Block-TT format. The proposed methods operate directly on the low-rank factor representation, preserve positivity by construction, and support adaptive rank refinement through tensor-network sweeps.
- We derive efficient tensor-network contraction schemes based on effective operators and recursively constructed environments, enabling scalable expected-value evaluation and local optimization.
- We extend the original conference framework with overlapping measurement operators and environment caching strategies, allowing efficient tomography of larger structured quantum systems.
- We provide an extensive numerical study comparing the proposed approaches with established low-rank tomography methods, demonstrating improved reconstruction accuracy, rank adaptivity, and computational efficiency across a range of problem settings.

1.3. Outline. Section 2 reviews the fundamentals of QST and TT representations, introducing the definitions used throughout the paper. Section 3 presents our main contributions, including a compressed sensing approach for low-rank QST in the proposed block-TT format in Subsection 3.2. Section 4 presents supporting numerical experiments and Section 5 concludes the paper.

2. Preliminaries. This section establishes the notation and reviews the concepts from low-rank QST and TT representations that are needed for the derivation of the Block-TT tomography algorithms in Section 3.

2.1. Notation. We use lowercase, bold lowercase, bold uppercase, bold uppercase with an underline, and calligraphic letters to denote scalars, vectors, matrices, block matrices, and tensors, respectively; that is, $x, \mathbf{x}, \mathbf{X}, \underline{\mathbf{X}}$ and $\mathcal{X}$. The trace of a

matrix $\mathbf{X}$ is denoted by $\mathrm{Tr}(\mathbf{X})$. The Frobenius norm and the nuclear norm (trace norm) are denoted by $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_\star$, respectively. The Kronecker delta is defined by $\delta_{ij}=1$ for $i=j$ and 0 otherwise. A state vector is denoted by $\boldsymbol{\psi}$, a density matrix by $\boldsymbol{\rho}$, and a d-dimensional Hilbert space by $\mathcal{H}_d$. The space of density matrices is denoted by $\mathcal{S}$, where $\mathcal{S} = \{\boldsymbol{\rho} \in \mathbb{C}^{D\times D} \mid \boldsymbol{\rho} = \boldsymbol{\rho}^{\mathrm{H}},\ \boldsymbol{\rho} \succeq 0,\ \mathrm{Tr}(\boldsymbol{\rho}) = 1\}$. The outer product $\mathcal{Z} = \mathbf{X} \otimes \mathbf{Y} \in \mathbb{C}^{I_1\times I_2\times J_1\times J_2}$ of two matrices $\mathbf{X} \in \mathbb{C}^{I_1\times I_2}$ and $\mathbf{Y} \in \mathbb{C}^{J_1\times J_2}$ satisfies $z_{ijkl} = x_{ij}y_{kl}$ and the Kronecker product is denoted by $\mathbf{Z} = \mathbf{X}\otimes\mathbf{Y} \in \mathbb{C}^{I_1J_1\times I_2J_2}$. The k-fold Kronecker product of $\mathbf{X}$ is denoted by $\mathbf{X}^{\otimes k} = \mathbf{X}\otimes\cdots\otimes\mathbf{X} \in \mathbb{C}^{I_1^k\times I_2^k}$. The *order* of a tensor $\mathcal{X} \in \mathbb{C}^{I_1\times I_2\times\cdots\times I_N}$ is the number of its *modes,* which corresponds to the number of free edges in the tensor network diagram, as shown in Figure 2.1. A specific entry of an Nth-order tensor is denoted by $x_{i_1i_2\cdots i_N} = \mathcal{X}(i_1, i_2, \ldots, i_N)$. The nth matrix unfolding of an Nth-order tensor is the matrix $\mathbf{X}_{[1,\ldots,n;n+1,\ldots,N]} \in \mathbb{C}^{I_1\cdots I_n\times I_{n+1}\cdots I_N}$ obtained by reshaping the tensor such that the first n modes form the row index and the remaining modes form the column index. The ordering of the indices determines how the tensor entries are stacked in the resulting matrix. For a tensor $\mathcal{X} \in \mathbb{C}^{R_1\times I_1\times J_1\times R_2}$ and its equivalent block-matrix representation $\underline{\mathbf{X}} = [\mathbf{X}_{r_1r_2}] \in \mathbb{C}^{R_1\cdot I_1\times J_1\cdot R_2}$, with blocks $\mathbf{X}_{r_1r_2} \in \mathbb{C}^{I_1\times J_1}$, we use $\mathbf{X}_{[1,2;3,4]} \in \mathbb{C}^{R_1I_1\times J_1R_2}$ for the corresponding matrix unfolding. The use of " $\cdot$ " in the dimension expressions indicates a block-matrix structure, with external indices defining the block layout and internal dimensions specifying the block size. For a block matrix $\underline{\mathbf{X}} \in \mathbb{C}^{R_1\cdot I_1\times J_1\cdot R_2}$, define the *left* and *right unfoldings* as $\mathbf{X}_{\mathrm{L}} = \mathbf{X}_{[1,2,3;4]} \in \mathbb{C}^{R_1I_1J_1\times R_2}$ and $\mathbf{X}_{\mathrm{R}} = \mathbf{X}_{[1;2,3,4]} \in \mathbb{C}^{R_1\times I_1J_1R_2}$. The block matrix is called *left-* or *right-orthogonal* if $\mathbf{X}_{\mathrm{L}}^{\mathrm{H}}\mathbf{X}_{\mathrm{L}} = \mathbf{I}_{R_2}$ or $\mathbf{X}_{\mathrm{R}}\mathbf{X}_{\mathrm{R}}^{\mathrm{H}} = \mathbf{I}_{R_1}$, respectively. The blockwise conjugate transpose of $\underline{\mathbf{X}} \in \mathbb{C}^{R_1\cdot I_1\times J_1\cdot R_2}$ is denoted by $\underline{\mathbf{X}}^{\mathrm{H}} \in \mathbb{C}^{R_1\cdot J_1\times I_1\cdot R_2}$ (see [30] for details). $\mathbf{X}_{\mathrm{TT}}$ denotes the matrix TT (TTM) decomposition of a matrix $\mathbf{X} \in \mathbb{C}^{\prod_{n=1}^N I_n\times\prod_{n=1}^N J_n}$; see (2.4) for a detailed definition. $\underline{\mathbf{X}}^{(<n)}$ and $\underline{\mathbf{X}}^{(>n)}$ denote the left and right-interface matrices, respectively, and $\mathbf{X}_{\neq n}$ the frame matrix; see Subsection 2.3.

DEFINITION 2.1 (Contraction). *A contraction of two tensors $\mathcal{X} \in \mathbb{C}^{I_1\times I_2\times\cdots\times I_N}$ and $\mathcal{Y} \in \mathbb{C}^{J_1\times J_2\times\cdots\times J_M}$, with common modes $I_n = J_m = K$, yields an $(N+M-2)$-order tensor $\mathcal{Z} = \mathcal{X} \bullet_n^m \mathcal{Y}$, the entries of which are given by:*

$$z_{i_1\ldots i_{n-1}i_{n+1}\ldots i_N j_1\ldots j_{m-1}j_{m+1}\ldots j_M} = \sum_{k=1}^{K} x_{i_1\ldots i_{n-1}ki_{n+1}\ldots i_N} y_{j_1\ldots j_{m-1}kj_{m+1}\ldots j_M}.$$

This is a fundamental operation in tensor networks and can be viewed as a multilinear extension of matrix–matrix multiplication.

DEFINITION 2.2 (Strong Kronecker product). *Let $\underline{\mathbf{X}} = [\mathbf{X}_{r_1r_2}] \in \mathbb{C}^{R_1\cdot I_1\times J_1\cdot R_2}$ and $\underline{\mathbf{Y}} = [\mathbf{Y}_{r_2r_3}] \in \mathbb{C}^{R_2\cdot I_2\times J_2\cdot R_3}$ be block matrices with $\mathbf{X}_{r_1r_2} \in \mathbb{C}^{I_1\times J_1}$ and $\mathbf{Y}_{r_2r_3} \in \mathbb{C}^{I_2\times J_2}$. The strong Kronecker product is defined by $\underline{\mathbf{Z}} = \underline{\mathbf{X}} \boxtimes \underline{\mathbf{Y}} \in \mathbb{C}^{R_1\cdot(I_1I_2)\times(J_1J_2)\cdot R_3}$ with blocks [6, 30]*

$$\mathbf{Z}_{r_1r_3} = \sum_{r_2=1}^{R_2} \mathbf{X}_{r_1r_2} \otimes \mathbf{Y}_{r_2r_3}.$$

This operation is analogous to block matrix multiplication, but with Kronecker products instead of matrix products.

DEFINITION 2.3 (Core (C) product). *Let $\underline{\mathbf{X}} = [\mathbf{X}_{r_1r_2}] \in \mathbb{C}^{R_1\cdot I_1\times J_1\cdot R_2}$ and $\underline{\mathbf{Y}} = [\mathbf{Y}_{r_3r_4}] \in \mathbb{C}^{R_3\cdot J_1\times J_2\cdot R_4}$ be block matrices with $\mathbf{X}_{r_1r_2} \in \mathbb{C}^{I_1\times J_1}$ and $\mathbf{Y}_{r_3r_4} \in \mathbb{C}^{J_1\times J_2}$.*

The C-product is defined by $\underline{\mathbf{Z}} = \underline{\mathbf{X}} \,|\bullet|\, \underline{\mathbf{Y}} \in \mathbb{C}^{(R_1R_3)\cdot I_1 \times J_2 \cdot (R_2R_4)}$ *with blocks [6, 30]*

$$\mathbf{Z}_{(r_1r_3)(r_2r_4)} = \mathbf{X}_{r_1r_2}\mathbf{Y}_{r_3r_4}.$$

This operation follows a standard Kronecker layout across the outer block indices, while combining the inner sub-blocks via standard matrix multiplication.

For ease of understanding, we use tensor network diagrams to visualize tensors and their operations; Figure 2.1 shows a few basic examples.

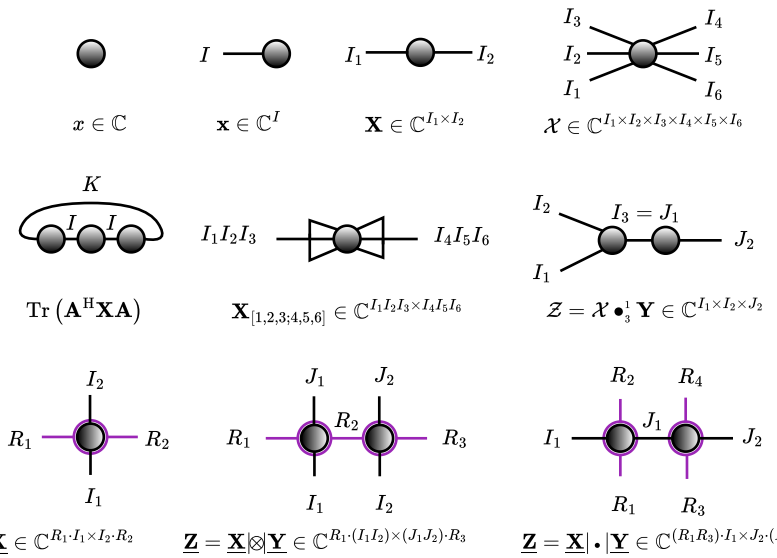

Fig. 2.1: Basic tensor network diagrams. In the bottom row, double-ring nodes represent block matrices with purple block-layout edges and black inner-matrix edges; block matrices are linked by a purple edge in the strong Kronecker product and a black edge in the C-product.

2.2. Quantum states, measurements and tomography. This section reviews quantum mechanical concepts like state representation, measurement operators, and existing low-rank QST approaches that motivate the Block-TT approach presented in Section 3.

State representation. A *qubit* (quantum bit) is the fundamental unit of quantum information. Unlike a classical bit, which takes a definite value 0 or 1, a qubit can exist in a superposition of both. Any pure qubit state can be represented as a unit vector in a two-dimensional Hilbert space as $|\psi\rangle = a|\mathbf{0}\rangle + b|\mathbf{1}\rangle$, where $|\mathbf{0}\rangle$ and $|\mathbf{1}\rangle$ form an orthonormal basis (written in Dirac notation) representing two distinct physical states of the qubit. The coefficients $a, b \in \mathbb{C}$ are probability amplitudes that specify the contribution of ψ along the basis directions $|\mathbf{0}\rangle$ and $|\mathbf{1}\rangle$. When a measurement is performed in this basis, the state collapses to either $|\mathbf{0}\rangle$ or $|\mathbf{1}\rangle$. However, in a random experiment, the probability of obtaining $|\mathbf{0}\rangle$ is $|a|^2$, and the probability of obtaining $|\mathbf{1}\rangle$ is $|b|^2$, according to the Born rule. The normalization condition $|a|^2 + |b|^2 = 1$ ensures that the total probability sums to one. Equivalently, the state vector can be represented as, $\psi = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathcal{H}_2 := \mathbb{C}^2$. A *qudit* is a d-dimensional extension of

a qubit, with pure states given by unit vectors in $\mathbb{C}^d$. A pure state of an N-qudit system is described by unit vectors of a Hilbert space formed by the tensor product of constituent Hilbert spaces. For more details, we refer to [44, 36, 43, 58, 49].

Thus far, we have only discussed pure states; a quantum system may also exist in a probabilistic mixture of pure states, in which the system occupies one of several $\boldsymbol{\psi}_k$ with associated probabilities α_k. That is, the system is described by the statistical ensemble $\{\alpha_k, \boldsymbol{\psi}_k\}$. A more general description of the (mixed) state is given by a statistical operator (density matrix) $\boldsymbol{\rho} = \sum_k \alpha_k \boldsymbol{\psi}_k \boldsymbol{\psi}_k^*$, where $\alpha_k \geq 0$, $\sum_k \alpha_k = 1$, and $\boldsymbol{\psi}_k$ are the (pure) state vectors. Following standard terminology in quantum information theory, we use "state" and "density matrix" in the rest of the paper interchangeably.

2.2.1. Quantum measurements. According to the basic postulates of quantum mechanics, observable (measurable) quantities of a quantum system are represented by Hermitian operators. Any such operator $\mathbf{M} = \mathbf{M}^{\mathrm{H}}$ admits a spectral decomposition $\mathbf{M} = \sum_k \lambda_k \mathbf{P}_k$, where the λ_k are real eigenvalues representing the possible outcomes of a measurement, and $\mathbf{P}_k$ are orthogonal projectors onto the corresponding eigenspaces, satisfying $\mathbf{P}_k \mathbf{P}_l = \delta_{kl} \mathbf{P}_k$ (orthogonality) and $\sum_k \mathbf{P}_k = \mathbf{I}$ (completeness) [44, 36, 43]. The probability p_{λ_k} of obtaining outcome λ_k is given by $\mathrm{Tr}\,(\boldsymbol{\rho}\mathbf{P}_k)$, and the overall expectation value is $\mathrm{Tr}\,(\boldsymbol{\rho}\mathbf{M})$. The completeness condition ensures that the measurement accounts for all possible outcomes and produces a properly normalized probability distribution[1]. A measurement defined by orthogonal projectors $\{\mathbf{P}_k\}$ satisfying completeness, and inducing a probability measure on outcomes, is commonly called a projective-valued measure (PVM) [44, 36].

Notice that, in a projective measurement framework, the number of outcomes is limited by the Hilbert space dimension, as there can be at most that many mutually orthogonal projectors. This limitation is addressed by a more general measurement framework, known as the positive operator-valued measure (POVM), which relaxes the orthogonality and projection requirements of PVM elements, replacing them with the weaker conditions of positivity and completeness, while still encoding the valid and meaningful probabilistic nature of the measurement outcomes [36, 43]. That is, a POVM is a collection of K measurement operators $\{\mathbf{E}_k\}_{k=1}^K$ satisfying $\mathbf{E}_k \succcurlyeq 0$ (positivity) and $\sum_{k=1}^K \mathbf{E}_k = \mathbf{I}$ (completeness). As a consequence, a POVM can have more elements than the dimension of the Hilbert space they act in. Each element $\mathbf{E}_k$ in the POVM is associated with a possible outcome of a quantum measurement. The probability p_k of getting the outcome with label k is given by $\mathrm{Tr}\,(\boldsymbol{\rho}\mathbf{E}_k)$, so a single-shot measurement yields label k at random with this probability. Consequently, estimating all $\{p_k\}$ requires multiple shots on identically prepared copies of the state, from which the probabilities are inferred via empirical frequencies [44, 49].

A POVM is *informationally complete* if its elements contain enough information to uniquely infer the state from the outcomes. Otherwise, one can combine multiple POVMs—say L of them—so that the union of the resulting $M = KL$ elements is informationally complete. A notable difference between POVMs and PVMs lies in their ability to be informationally complete. For a generic mixed state, the density matrix has a number of independent parameters that scales quadratically with the Hilbert space dimension. A single PVM, however, yields at most a number of independent outcomes equal to that dimension and is therefore never informationally complete. In contrast, a single POVM is not subject to this restriction; a single POVM can

[1] That is $\sum_k p_{\lambda_k} = \sum_k \mathrm{Tr}(\boldsymbol{\rho}\mathbf{P}_k) = \mathrm{Tr}(\boldsymbol{\rho} \sum_k \mathbf{P}_k) = \mathrm{Tr}(\boldsymbol{\rho}\mathbf{I}) = \mathrm{Tr}(\boldsymbol{\rho}) = 1$.

therefore be informationally complete. See [44, 36, 43, 49] for more details.

The *Pauli spin matrices* for a single qubit are denoted by $\mathbf{I}_2, \boldsymbol{\sigma}_x := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \boldsymbol{\sigma}_y := \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \boldsymbol{\sigma}_z := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. They form an informationally complete set for a single-qubit system. Notice that this set of matrices is not a POVM itself, but measuring each of these observables is described by a two-outcome POVM of the form $\{\frac{\mathbf{I}_2+\boldsymbol{\sigma}_n}{2}, \frac{\mathbf{I}_2-\boldsymbol{\sigma}_n}{2}\}$, where $\boldsymbol{\sigma}_n \in \{\mathbf{I}_2, \boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y, \boldsymbol{\sigma}_z\}$. Pauli matrices are commonly used as measurement observables in single-qubit systems. For an N-qubit system, local[2] measurement observables can be formed as Kronecker products of single-qubit Pauli matrices, $\mathbf{E}_m := \otimes_{n=1}^{N} \boldsymbol{\sigma}_n$, where $\boldsymbol{\sigma}_n$ denotes the Pauli operator acting on qubit n.

Another commonly used local (i.e., single-qubit) measurement is the symmetric informationally complete (SIC) POVM [50], given by

$$\left\{ \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{6} & \frac{\sqrt{2}}{6} \\ \frac{\sqrt{2}}{6} & \frac{1}{6} \end{bmatrix}, \begin{bmatrix} \frac{1}{6} & \frac{\sqrt{2}}{6} e^{-i\frac{2\pi}{3}} \\ \frac{\sqrt{2}}{6} e^{i\frac{2\pi}{3}} & \frac{1}{6} \end{bmatrix}, \begin{bmatrix} \frac{1}{6} & \frac{\sqrt{2}}{6} e^{-i\frac{4\pi}{3}} \\ \frac{\sqrt{2}}{6} e^{i\frac{4\pi}{3}} & \frac{1}{6} \end{bmatrix} \right\}.$$

Low-rank QST. QST can be formulated as an inverse problem. Consider an N-qudit system defined on the Hilbert space $\mathcal{H}_D := (\mathbb{C}^d)^{\otimes N}$, with total dimension $D = d^N$. The state of the system is described by a density matrix $\boldsymbol{\rho} \in \mathbb{C}^{D\times D}$, belonging to $\mathcal{S}$, the set of Hermitian, PSD matrices with unit trace. In the special case of qubits, we have $d = 2$. Suppose we measure an informationally complete set of POVM elements, obtaining (possibly noisy) measurement statistics $\{\mathbf{E}_m, y_m\}_{m=1}^{M}$. The goal of QST is to reconstruct a density matrix $\hat{\boldsymbol{\rho}}$ consistent with the observations, i.e., $y_m \approx \mathrm{Tr}(\hat{\boldsymbol{\rho}}\mathbf{E}_m)$, for $m = 1, 2, \ldots, M$. This recovery problem is challenging because (1) $\hat{\boldsymbol{\rho}}$ must satisfy the physicality constraints, i.e., $\hat{\boldsymbol{\rho}} \in \mathcal{S}$, and (2) the number of measurement settings required to ensure unique recovery of a generic, unstructured density matrix scales as $\mathcal{O}(D^2)$ [62]. However, if the underlying density matrix is low-rank, it has been shown that significantly fewer measurements suffice to estimate it. Under Pauli measurements, the rank-R matrix has been shown to be uniquely estimated using $\mathcal{O}(RD\log^2 D)$ measurement settings, with high probability [16, 32, 31]. Under structured measurements, $\mathcal{O}(RD)$ settings may suffice [2, 57, 55]. The literature outlines the following two approaches to employing the low-rank constraints.

Rank minimization. In this method, the low-rank constraint is enforced implicitly by formulating a rank-minimization objective (with nuclear norm relaxation of the rank), under measurements and physicality constraints. Mathematically [16],

$$\min_{\hat{\boldsymbol{\rho}} \in \mathbb{C}^{D\times D}} \|\hat{\boldsymbol{\rho}}\|_\star \quad \text{s.t.} \quad \|\mathbf{y} - \mathcal{M}(\hat{\boldsymbol{\rho}})\|_2 \le \epsilon, \text{ and } \hat{\boldsymbol{\rho}} \succeq 0, \tag{2.1}$$

where $\mathbf{y} \in \mathbb{R}^M$ are the measurement outcomes, and the measurement (sensing) map $\mathcal{M}$ is $(\mathcal{M}(\boldsymbol{\rho}))_m = \mathrm{Tr}(\mathbf{E}_m\boldsymbol{\rho})$. To ensure $\hat{\boldsymbol{\rho}} \in \mathcal{S}$, one generally normalizes the solution as $\hat{\boldsymbol{\rho}} := \hat{\boldsymbol{\rho}}/\mathrm{Tr}(\hat{\boldsymbol{\rho}})$; for noiseless, trace-preserving measurements, the unit-trace condition is automatically satisfied under the recovery conditions [21]. The nuclear norm relaxation is the tightest convex surrogate of the rank function; directly minimizing the rank leads to a non-convex and NP-hard problem [60, 13]. This relaxation enables an approximate yet computationally more efficient solution. The problem (2.1) can be formulated as a semidefinite program (SDP) [12, 13], and when N is small, the SDP can be readily solved using software such as SDPT3 or CVXPY [9]. For mid- to large-scale systems, optimizing over a full $(D \times D)$ density matrix becomes infeasible.

[2] That is, acting independently on individual qubits.

Error minimization. This approach explicitly promotes low-rank structure by parametrizing the density matrix using a fixed, low-rank Burer–Monteiro (Cholesky-like) factorization [4]. While this parameterization renders the problem non-convex, it significantly improves computational efficiency by restricting optimization to low-rank factors. Mathematically [4],

$$\min_{\mathbf{A}\in\mathbb{C}^{D\times R}} \|\mathbf{y}-\mathcal{M}(\hat{\boldsymbol{\rho}})\|_2^2 \quad \text{with } \hat{\boldsymbol{\rho}}=\mathbf{A}\mathbf{A}^{\mathrm{H}},\ \text{s.t. } \mathrm{Tr}(\hat{\boldsymbol{\rho}})=1. \tag{2.2}$$

Here, the flexibility to choose a differentiable objective function allows for the use of gradient-based optimization approaches (e.g., first- and/or second-order methods), yielding an efficient approximate solution to QST. Moreover, the Hermitian and PSD constraints are automatically satisfied by construction. In general, the optimization problems arising from this parameterization may admit local solutions[3]. However, with mild regularity conditions and suitable initialization, this approach becomes effective for many applications, including QST [3, 4, 21, 27, 24]. In practice, the unit-trace requirement is often relaxed to the convex constraint $\mathrm{Tr}(\hat{\boldsymbol{\rho}})\le 1 \iff \|\mathbf{A}\|_F^2\le 1$, to further improve scalability [3, 27]. While the parameters of $\mathbf{A}$ remain exponential in N, our approach generalizes the Burer–Monteiro factorization to the Block-TT format, breaking this exponential dependence.

2.3. Tensor train decomposition. A TT decomposition of an Nth-order tensor $\mathcal{X}\in\mathbb{C}^{I_1\times I_2\times\cdots\times I_N}$ corresponds to a contraction of a sequence of third-order tensors (TT-cores) $\mathcal{X}^{(n)}\in\mathbb{C}^{R_{n-1}\times I_n\times R_n}(1\le n\le N)$ with $R_0=R_N=1$ such that each entry of $\mathcal{X}$ can be expressed as the sequence of matrix products [46, 42]:

$$x_{i_1 i_2\cdots i_{N-1} i_N}=\mathbf{X}^{(1)}_{:i_1:}\mathbf{X}^{(2)}_{:i_2:}\cdots\mathbf{X}^{(N-1)}_{:i_{N-1}:}\mathbf{X}^{(N)}_{:i_N:}, \tag{2.3}$$

where the matrix $\mathbf{X}^{(n)}_{:i:}\in\mathbb{C}^{R_{n-1}\times R_n}$ is the ith mode-2 slice of the TT-core $\mathcal{X}^{(n)}$. The tuple of minimal integers $(R_0,\ldots,R_N)$ for which equality in Equation (2.3) holds is the TT-rank of $\mathcal{X}$.

A dense tensor $\mathcal{X}$ can be decomposed into a TT format via the TT-SVD algorithm, which successively reshapes the tensor into unfolding matrices with one isolated mode and computes truncated SVDs from left to right. At each step, the left singular vectors form a TT-core, while the remaining factor is reshaped and propagated to the subsequent decomposition step [42]. Due to its sequential nature, TT-SVD is difficult to parallelize; alternatively, Parallel-TTSVD computes truncated SVDs of all nth unfoldings independently and then constructs the TT-cores from them [53].

TT representations for vectors and matrices. Large-scale vectors and matrices can be represented in TT format after tensorization. Specifically, a vector $\mathbf{x}\in\mathbb{C}^I$ is reshaped into a tensor $\mathcal{X}\in\mathbb{C}^{I_1\times\cdots\times I_N}$, while a matrix $\mathbf{X}\in\mathbb{C}^{I\times J}$ is reshaped into a tensor $\mathcal{X}\in\mathbb{C}^{(I_1J_1)\times\cdots\times(I_NJ_N)}$, where $I=\prod_{n=1}^N I_n$ and $J=\prod_{n=1}^N J_n$. The corresponding TT-cores are third-order tensors $\mathcal{X}^{(n)}\in\mathbb{C}^{R_{n-1}\times I_n\times R_n}$ for vectors and fourth-order tensors $\mathcal{X}^{(n)}\in\mathbb{C}^{R_{n-1}\times I_n\times J_n\times R_n}$ for matrices, obtained by splitting the "long mode" I_nJ_n into $I_n\times J_n$ [42, 40]. In this paper, we adopt an equivalent TT representation using the strong Kronecker product, together with the C-product to describe tensor network contractions. This notation provides a more intuitive framework for visualizing operations on tensors obtained by tensorizing vectors and matrices.

[3]Without any additional constraints beyond the PSD constraint, one can decompose $\boldsymbol{\rho}$ as $\boldsymbol{\rho}=\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathrm{H}}$, where $\hat{\mathbf{A}}=\mathbf{A}\mathbf{L}$ for any unitary $\mathbf{L}\in\mathbb{C}^{R\times R}$ such that $\mathbf{L}\mathbf{L}^{\mathrm{H}}=\mathbf{I}$.

In the matrix case, a matrix TT (TTM [40]), mathematically equivalent to the MPO, can also be represented as a sequential strong Kronecker product of N block matrices,

$$\mathbf{X}_{\mathrm{TT}} = \underline{\mathbf{X}}^{(1)} \boxtimes \underline{\mathbf{X}}^{(2)} \boxtimes \cdots \boxtimes \underline{\mathbf{X}}^{(N)} \in \mathbb{C}^{\prod_{n=1}^{N} I_n \times \prod_{n=1}^{N} J_n}, \tag{2.4}$$

where each core $\underline{\mathbf{X}}^{(n)} \in \mathbb{C}^{R_{n-1} \cdot I_n \times J_n \cdot R_n}$ consists of blocks $\mathbf{X}^{(n)}_{r_{n-1}::r_n} \in \mathbb{C}^{I_n \times J_n}$:

$$\underline{\mathbf{X}}^{(n)} = \begin{bmatrix} \mathbf{X}^{(n)}_{1::1} & \cdots & \mathbf{X}^{(n)}_{1::R_n} \\ \vdots & \ddots & \vdots \\ \mathbf{X}^{(n)}_{R_{n-1}::1} & \cdots & \mathbf{X}^{(n)}_{R_{n-1}::R_n} \end{bmatrix}.$$

The corresponding tensor entries satisfy[4]

$$x_{i_1 j_1 i_2 j_2 \ldots i_N j_N} = \mathbf{X}^{(1)}_{:i_1 j_1 :} \mathbf{X}^{(2)}_{:i_2 j_2 :} \cdots \mathbf{X}^{(N)}_{:i_N j_N :}. \tag{2.5}$$

Note that $\underline{\mathbf{X}}^{(n)} \in \mathbb{C}^{R_{n-1} \cdot I_n \times J_n \cdot R_n}$ is referred to as a *core*, while its corresponding tensor in the TT decomposition $\mathcal{X}^{(n)} \in \mathbb{C}^{R_{n-1} \times I_n \times J_n \times R_n}$ is referred to as a *TT-core*. The TT-rank $(R_0, \ldots, R_N)$ corresponds to the number of blocks in $\underline{\mathbf{X}}^{(n)}$, with $R_0 = R_N = 1$. Beyond its standard view as a tuple of unfolding matrix ranks quantifying row and column independence, the TT-rank explicitly captures the internal structural complexity of the strong Kronecker product representation.

In the vector case $J = 1$ (and hence $J_n = 1$, $\forall n$), each block of $\underline{\mathbf{X}}^{(n)}$ reduces to a vector of size I_n, yielding the standard TT/MPS representation.

A Block–n–TT is a TTM where all cores consist of vector-valued blocks except for a single matrix-valued block core at position n ($J_i = 1$ for $i \neq n$, $J_n = K$) [47, 10, 25, 29]. As shown in Figure 2.2, it is equivalent to a collection of TT vectors differing only in their nth core.

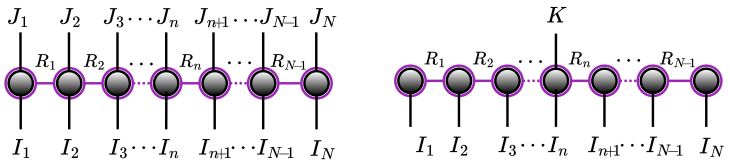

Fig. 2.2: (Left) TTM representation of a matrix $\mathbf{X} \in \mathbb{C}^{I_1 I_2 \cdots I_N \times J_1 J_2 \cdots J_N}$. (Right) Block–$n$–TT representation of a tall matrix $\mathbf{X} \in \mathbb{C}^{I_1 I_2 \cdots I_N \times K}$, or equivalently, as a collection of K TT vectors that share all cores, except for the nth core.

A TTM can also be represented using the shorthand notation

$$\mathbf{X}_{\mathrm{TT}} = \underline{\mathbf{X}}^{(<n)} \boxtimes \underline{\mathbf{X}}^{(n)} \boxtimes \underline{\mathbf{X}}^{(>n)} = \underline{\mathbf{X}}^{(<n)} \boxtimes \underline{\mathbf{X}}^{(n,n+1)} \boxtimes \underline{\mathbf{X}}^{(>n+1)}, \tag{2.6}$$

where the left-interface $\underline{\mathbf{X}}^{(<n)} = \underline{\mathbf{X}}^{(1)} \boxtimes \cdots \boxtimes \underline{\mathbf{X}}^{(n-1)} \in \mathbb{C}^{R_0 \cdot \left(\prod_{i=1}^{(n-1)} I_i\right) \times \left(\prod_{i=1}^{(n-1)} J_i\right) \cdot R_{n-1}}$, the right-interface $\underline{\mathbf{X}}^{(>n)} = \underline{\mathbf{X}}^{(n+1)} \boxtimes \cdots \boxtimes \underline{\mathbf{X}}^{(N)} \in \mathbb{C}^{R_n \cdot (\prod_{i=n+1}^{N} I_i) \times (\prod_{i=n+1}^{N} J_i) \cdot R_N}$ and the merged core $\underline{\mathbf{X}}^{(n,n+1)} = \underline{\mathbf{X}}^{(n)} \boxtimes \underline{\mathbf{X}}^{(n+1)} \in \mathbb{C}^{R_{n-1} \cdot (I_n I_{n+1}) \times (J_n J_{n+1}) \cdot R_{n+1}}$. The merged

[4] The strong Kronecker product representation formally yields a matrix; here, entries are understood via the associated tensorized representation.

core is commonly referred to as a "supercore" or a "two-site tensor" in the DMRG literature [63, 52, 23, 41, 18, 25, 29].

If two TTMs $\mathbf{X}_{\text{TT}} = \underline{\mathbf{X}}^{(1)} \circledast \cdots \circledast \underline{\mathbf{X}}^{(N)}$ and $\mathbf{Y}_{\text{TT}} = \underline{\mathbf{Y}}^{(1)} \circledast \cdots \circledast \underline{\mathbf{Y}}^{(N)}$ have the same mode sizes, their linear combination $\mathbf{Z}_{\text{TT}} = \alpha\mathbf{X}_{\text{TT}} + \beta\mathbf{Y}_{\text{TT}}$, can be expressed as

$$\mathbf{Z}_{\text{TT}} = \begin{bmatrix}\underline{\mathbf{X}}^{(1)} & \underline{\mathbf{Y}}^{(1)}\end{bmatrix} \circledast \begin{bmatrix}\underline{\mathbf{X}}^{(2)} & \mathbf{0} \\ \mathbf{0} & \underline{\mathbf{Y}}^{(2)}\end{bmatrix} \circledast \cdots \circledast \begin{bmatrix}\underline{\mathbf{X}}^{(N-1)} & \mathbf{0} \\ \mathbf{0} & \underline{\mathbf{Y}}^{(N-1)}\end{bmatrix} \circledast \begin{bmatrix}\alpha\underline{\mathbf{X}}^{(N)} \\ \beta\underline{\mathbf{Y}}^{(N)}\end{bmatrix}.$$

The multiplication $\mathbf{Z}_{\text{TT}} = \mathbf{X}_{\text{TT}}\mathbf{Y}_{\text{TT}}$, with $\mathbf{X}_{\text{TT}}$ and $\mathbf{Y}_{\text{TT}}$ having compatible mode sizes, can be compactly written as

$$\mathbf{Z}_{\text{TT}} = \left(\underline{\mathbf{X}}^{(1)} \,|\bullet|\, \underline{\mathbf{Y}}^{(1)}\right) \circledast \cdots \circledast \left(\underline{\mathbf{X}}^{(N)} \,|\bullet|\, \underline{\mathbf{Y}}^{(N)}\right).$$

Notice that TTM addition concatenates cores block-diagonally (direct sum), upper-bounding the ranks by the sum of the corresponding core ranks, whereas TTM multiplication combines cores via C-products, upper-bounding the ranks by their product. In practice, one performs the TT-rounding operation [42] to keep the ranks minimal.

DEFINITION 2.4 (Frame equation for Block-TT). *A TT(M) representation is multilinear with respect to its cores. Consequently, the Block–n–TT representation ($J_i = 1$ for $i \neq n$, $J_n = K$) is linear with respect to the nth core. The same property holds for a merged two-site core. It can be expressed using the matrix frame equation, as defined in [25, 10, 29]:*

$$\mathbf{X}_{\text{TT}} = \mathbf{X}_{\neq n}\, \mathbf{X}^{(n)}_{[1,2,4;3]} = \mathbf{X}_{\neq n,n+1}\, \mathbf{X}^{(n,n+1)}_{[1,2,4;3]} \in \mathbb{C}^{I_1 I_2 \cdots I_N \times K}. \tag{2.7}$$

The corresponding frame matrices are constructed from the left and right interfaces as $\mathbf{X}_{\neq n} = \mathbf{X}_{\text{L}}^{(<n)} \otimes \mathbf{I}_{I_n} \otimes \mathbf{X}_{\text{R}}^{(>n)\text{H}} \in \mathbb{C}^{(\prod_{i=1}^{N} I_i)\times(R_{n-1}I_nR_n)}$, and $\mathbf{X}_{\neq n,n+1} = \mathbf{X}_{\text{L}}^{(<n)} \otimes \mathbf{I}_{I_n} \otimes \mathbf{I}_{I_{n+1}} \otimes \mathbf{X}_{\text{R}}^{(>n+1)\text{H}} \in \mathbb{C}^{(\prod_{i=1}^{N} I_iJ_i)\times(R_{n-1}I_nI_{n+1}R_{n+1})}$. Here, $\mathbf{X}^{(n)}_{[1,2,4;3]} \in \mathbb{C}^{R_{n-1}I_nR_n\times K}$ and $\mathbf{X}^{(n,n+1)}_{[1,2,4;3]} \in \mathbb{C}^{R_{n-1}I_nI_{n+1}R_{n+1}\times K}$ are the matrix unfoldings of the core $\underline{\mathbf{X}}^{(n)}$ and the merged two-site core $\underline{\mathbf{X}}^{(n,n+1)}$, respectively.

DEFINITION 2.5 (*n*–orthogonal form). *A TT(M) is in n–orthogonal form if all cores to the left of the nth core are left-orthogonal and all cores to its right are right-orthogonal. To obtain the n–orthogonality, one proceeds sequentially from the first (or last) core. At each step, a QR (or LQ) decomposition is applied to the left (or right) unfolding of the ith core. The orthogonal factor is reshaped back into a core, while the triangular factor is absorbed into the next (or previous) core. This procedure is repeated until the n–orthogonal form is achieved [42, 52, 64, 18, 25, 10]. Notice that n–orthogonality ensures orthogonality of the left- and right-interface matrices, i.e.,*

$$\mathbf{X}_{\text{L}}^{(<n)\text{H}}\mathbf{X}_{\text{L}}^{(<n)} = \mathbf{I}_{R_{n-1}}, \qquad \mathbf{X}_{\text{R}}^{(>n)}\mathbf{X}_{\text{R}}^{(>n)\text{H}} = \mathbf{I}_{R_n}.$$

Consequently, the frame matrices $\mathbf{X}_{\neq n}$ and $\mathbf{X}_{\neq n,n+1}$ are column-wise orthogonal. In the literature, 1-orthogonal and N-orthogonal forms are also referred to as right-orthogonal and left-orthogonal forms, respectively.

3. Block-TT approach to low-rank QST. This section presents our main contribution. We use an efficient extension of the Burer–Monteiro factorization to tensors to circumvent the curse of dimensionality in QST. More specifically, we show in Subsection 3.1 that the density matrix can be parameterized as a contraction of two Block-TT networks, thereby enforcing Hermiticity and positivity by construction. This parameterization enables a DMRG-like optimization scheme for QST of low-rank mixed states, which is developed in Subsection 3.2.

3.1. Block-TT for mixed low-rank states. We employ a representation that expresses a density matrix as a contraction of a Block–n–TT network with its Hermitian transpose, i.e.,

$$\boldsymbol{\rho}_{\mathsf{TT}} = \mathbf{A}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}} \in \mathbb{C}^{d^N \times d^N}, \quad \mathbf{A}_{\mathsf{TT}} = \underline{\mathbf{A}}^{(1)} \,|\otimes|\, \underline{\mathbf{A}}^{(2)} \,|\otimes|\, \cdots \,|\otimes|\, \underline{\mathbf{A}}^{(N)} \in \mathbb{C}^{d^N \times K}, \tag{3.1}$$

or entry-wise, we can write:

$$\rho_{\mathsf{TT}\, i_1 j_1 \ldots i_N j_N} = \boldsymbol{\rho}^{(1)}_{:i_1 j_1:} \boldsymbol{\rho}^{(2)}_{:i_2 j_2:} \cdots \boldsymbol{\rho}^{(N)}_{:i_N j_N:}, \quad \boldsymbol{\rho}^{(i)} = \underline{\mathbf{A}}^{(i)} \,|\bullet|\, \underline{\mathbf{A}}^{(i)\mathrm{H}}. \tag{3.2}$$

Here, $\mathbf{A}_{\mathsf{TT}} \in \mathbb{C}^{d^N \times K}$ bounds the matrix rank of $\boldsymbol{\rho}_{\mathsf{TT}}$ by K. Cores $\underline{\mathbf{A}}^{(i)} \in \mathbb{C}^{R_{i-1} \cdot d \times R_i}$ are vector-valued block cores for $i \neq n$, while the nth core is a matrix-valued block core, $\underline{\mathbf{A}}^{(n)} \in \mathbb{C}^{R_{n-1} \cdot d \times K \cdot R_n}$. Intuitively, this construction is analogous to the Burer–Monteiro factorization for matrices [4]. A tensor network diagram of this model is shown in Figure 3.1.

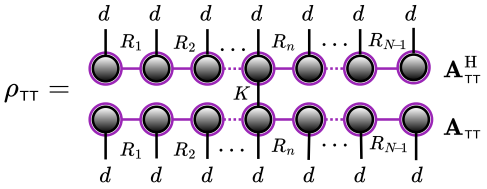

Fig. 3.1: Burer–Monteiro parameterization of a density matrix in Block–n–TT format, i.e., $\boldsymbol{\rho}_{\mathsf{TT}} = \mathbf{A}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}$, where $\mathbf{A}_{\mathsf{TT}} \in \mathbb{C}^{D \times K}$, $D = d^N$.

Physicality constraints. As noted earlier, a Block-TT can be interpreted as a collection of TT vectors sharing all cores except the nth core (i.e., the core carrying the block index K). The corresponding frame equation can be written in the following form (cf. Equation (2.7)): $\mathbf{A}_{\mathsf{TT}} = \mathbf{A}_{\neq n}\, \mathbf{A}^{(n)}_{[1,2,4;3]}$, where the reshaped nth core is $\mathbf{A}^{(n)}_{[1,2,4;3]} \in \mathbb{C}^{R_{n-1} d R_n \times K}$, and the frame matrix is given by $\mathbf{A}_{\neq n} = \mathbf{A}_{\mathsf{L}}^{(<n)} \otimes \mathbf{I}_d \otimes \mathbf{A}_{\mathsf{R}}^{(>n)\mathrm{H}} \in \mathbb{C}^{D \times R_{n-1} d R_n}$. Assume we work under n-orthogonality (see Definition 2.5). Then the spectrum of $\boldsymbol{\rho}_{\mathsf{TT}}$ satisfies

$$\sigma\big(\mathbf{A}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}\big) = \sigma\Big(\mathbf{A}_{\neq n}\, \mathbf{A}^{(n)}_{[1,2,4;3]} \mathbf{A}^{(n)\mathrm{H}}_{[1,2,4;3]}\, \mathbf{A}^{\mathrm{H}}_{\neq n}\Big) = \sigma\Big(\mathbf{A}_{\neq n}\, \boldsymbol{\rho}^{(n)}_{\mathsf{eff}}\, \mathbf{A}^{\mathrm{H}}_{\neq n}\Big) = \sigma\Big(\boldsymbol{\rho}^{(n)}_{\mathsf{eff}}\Big). \tag{3.3}$$

Therefore, the eigenvalues of the effective density matrix $\boldsymbol{\rho}^{(n)}_{\mathsf{eff}} = \mathbf{A}^{(n)}_{[1,2,4;3]} \mathbf{A}^{(n)\mathrm{H}}_{[1,2,4;3]} \in \mathbb{C}^{R_{n-1} d R_n \times R_{n-1} d R_n}$ coincide with the nonzero eigenvalues of $\boldsymbol{\rho}_{\mathsf{TT}}$. Consequently, the normalization constraints (unit-trace) also follow from local constraints, i.e., $\mathrm{Tr}(\boldsymbol{\rho}_{\mathsf{TT}}) = \mathrm{Tr}\Big(\boldsymbol{\rho}^{(n)}_{\mathsf{eff}}\Big) = 1$. Notice that the effective density matrix has a Gram structure and therefore guarantees that $\boldsymbol{\rho}_{\mathsf{TT}}$ is positive semidefinite by construction.

We observe that, under the n-orthogonality constraints, the structural properties of the global density matrix $\boldsymbol{\rho}_{\mathsf{TT}} = \mathbf{A}_{\neq n}\, \boldsymbol{\rho}^{(n)}_{\mathsf{eff}}\, \mathbf{A}^{\mathrm{H}}_{\neq n}$ are reflected in the effective density matrix. Owing to this renormalization-type (scale/transformation-invariant) structure, an SDP optimization problem posed on $\boldsymbol{\rho}_{\mathsf{TT}}$ can be reduced to a local problem on $\boldsymbol{\rho}^{(n)}_{\mathsf{eff}}$, which preserves the SDP structure provided that n-orthogonality holds and optimization constraints map linearly under frame projections. Note, however, that

we do not solve the SDP for $\boldsymbol{\rho}_{\mathsf{eff}}^{(n)}$ directly. Instead, we optimize over the low-rank factors $\mathbf{A}_{[1,2,4;3]}^{(n)}$, which results in a nonconvex optimization problem. After solving the local problem, the orthogonality center can be moved to the next (or previous) core, for instance, by shifting the core carrying the block index K and applying a QR (or LQ) step, as in Definition 2.5. See the rank adaptation and block-index sweeping in Subsection 3.2 for details.

We emphasize that the Block-TT parameterization of the density matrix is efficient only when the mixedness is low, i.e., when the density matrix can be expressed as a sum of a few pure states. This parametrization allows for representing a mixed state with asymptotically the same number of parameters as a TT/MPS. That is, the number of parameters is equal to $dR^2(\log_d D - 3) + dR^2K + 2dR$, assuming $R_n = R$.

3.2. Compressed sensing approach. We extend the error-minimization low-rank QST formulation (2.2) by parameterizing the density matrix via Block-TT network contractions. This approach offers two main advantages: logarithmic parameter scaling in the state dimension and fast expectation-value evaluation via tensor-network contractions. The low-rank Burer–Monteiro QST is formulated as:

$$\min_{\mathbf{A}_{\mathsf{TT}}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathcal{M}(\hat{\boldsymbol{\rho}}_{\mathsf{TT}})\|_2^2}_{=:f(\mathbf{A}_{\mathsf{TT}})}, \quad \text{s.t. } \|\mathbf{A}_{\mathsf{TT}}\|_{\mathrm{F}}^2 \le 1, \quad \text{with } \hat{\boldsymbol{\rho}}_{\mathsf{TT}} = \mathbf{A}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}. \tag{3.4}$$

The map $\mathcal{M}$ provides compressed (trace-based) measurements of the underlying density matrix, $(\mathcal{M}(\hat{\boldsymbol{\rho}}_{\mathsf{TT}}))_m = \mathrm{Tr}(\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}\mathbf{E}_m\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}})$. Each measurement yields the expectation value of an observable and generally depends on all density matrix entries.

DMRG-based approach to scalable QST. We focus on measurement operators that admit low-rank TTM representations, as is the case for many important classes of measurement operators, including Pauli observables and a broad range of structured operators that can be well approximated by sums of a relatively small number of Kronecker-product terms. Leveraging this structure, we first develop an efficient tensor-network contraction scheme for expectation-value evaluation, followed by a single-site DMRG optimization scheme for QST and a brief discussion of its two-site extension.

Consider measurement operators of the form,

$$\mathbf{E}_{\mathsf{TT}} = \underline{\mathbf{E}}^{(1)} \,|\otimes|\, \underline{\mathbf{E}}^{(2)} \,|\otimes|\, \cdots \,|\otimes|\, \underline{\mathbf{E}}^{(N)},$$

where the cores satisfy $\underline{\mathbf{E}}^{(n)} \in \mathbb{C}^{R_{n-1}^E \cdot d \times d \cdot R_n^E}$. There are various choices of observables; for example, for N-qubit systems ($d=2$), a common choice is the set of Pauli operators, which admit a TTM representation with TT-rank elements equal to one, i.e., $R_n^E = 1$, $\forall n$. In this case, the strong Kronecker product reduces to the standard one.

Efficient expectation values via tensor network contraction. The expectation value $\mathrm{Tr}(\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}\mathbf{E}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}})$ can be computed efficiently via corewise contractions:

$$\mathrm{Tr}\left(\left(\underline{\mathbf{A}}^{(1)\mathrm{H}} \,|\bullet|\, \underline{\mathbf{E}}^{(1)} \,|\bullet|\, \underline{\mathbf{A}}^{(1)}\right) |\otimes| \cdots |\otimes| \left(\underline{\mathbf{A}}^{(N)\mathrm{H}} \,|\bullet|\, \underline{\mathbf{E}}^{(N)} \,|\bullet|\, \underline{\mathbf{A}}^{(N)}\right) \right).$$

Figure 3.2 visualizes the tensor network diagram for this operation: the overall contraction between the Block-TT parameterized density operator and the (TTM) measurement operators (left), alongside its equivalent effective-operator form (right).

In practice, the effective operator $\mathbf{E}_{\text{eff}}^{(n)} = \mathbf{A}_{\neq n}^{\text{H}} \, \mathbf{E}_{\text{TT}} \, \mathbf{A}_{\neq n}$ is never explicitly formed; instead, it is constructed via the so-called *left and right environments*, $\mathcal{L}^{(<n)}$ and $\mathcal{R}^{(>n)}$, which are computed recursively as shown in (3.6). Once $\mathcal{L}^{(<n)} \in \mathbb{C}^{R_{n-1} \times R_{n-1}^E \times R_{n-1}}$ and $\mathcal{R}^{(>n)} \in \mathbb{C}^{R_n \times R_n^E \times R_n}$ are obtained, contracting the tensorized form $\mathcal{E}^{(n)} \in \mathbb{C}^{R_{n-1}^E \times d \times d \times R_n^E}$ of the local operator $\underline{\mathbf{E}}^{(n)}$ with these environments yields:

$$\mathcal{Z} = \left(\mathcal{E}^{(n)} \bullet_1^2 \mathcal{L}^{(<n)} \right) \bullet_3^2 \mathcal{R}^{(>n)}; \quad \mathbf{E}_{\text{eff}}^{(n)} = \mathbf{Z}_{[3,1,5\,;\,4,2,6]} \in \mathbb{C}^{R_{n-1} dR_n \times R_{n-1} dR_n}. \tag{3.5}$$

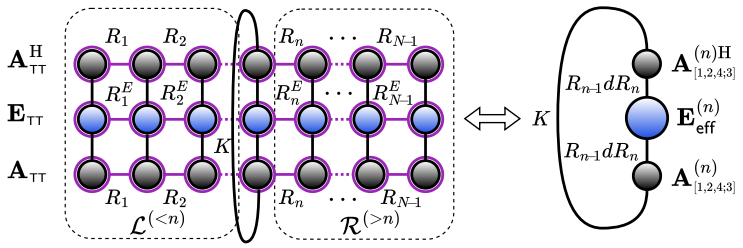

Fig. 3.2: (Left) Core-wise contraction of the tensor network representing the expectation value computation, i.e., $\text{Tr}(\mathbf{A}_{\text{TT}}^{\text{H}} \, \mathbf{E}_{\text{TT}} \, \mathbf{A}_{\text{TT}})$. (Right) Equivalent tensor network representation of the same operation in terms of an effective operator, i.e., $\text{Tr}(\mathbf{A}_{[1,2,4;3]}^{(n)\text{H}} \, \mathbf{E}_{\text{eff}}^{(n)} \, \mathbf{A}_{[1,2,4;3]}^{(n)})$, where $\mathbf{E}_{\text{eff}}^{(n)} = \mathbf{A}_{\neq n}^{\text{H}} \, \mathbf{E}_{\text{TT}} \, \mathbf{A}_{\neq n} \in \mathbb{C}^{R_{n-1} dR_n \times R_{n-1} dR_n}$.

Optimal contraction strategy. The tensor network in Figure 3.2 is contracted efficiently by recursively constructing the left and right environments, with boundary conditions $\mathcal{L}^{(<1)} = 1$ and $\mathcal{R}^{(>N)} = 1$. As illustrated in Figure 3.3, $\mathcal{L}^{(<n)}$ is computed by successively contracting $\mathcal{L}^{(<n-1)}$ with $\underline{\mathbf{A}}^{(n-1)\text{H}}$, $\underline{\mathbf{E}}^{(n-1)}$, and $\underline{\mathbf{A}}^{(n-1)}$; the right environments are obtained analogously. Specifically, let

$$\mathbf{Z}^{(i)} = \underline{\mathbf{A}}^{(i)\text{H}} \, | \bullet | \, \underline{\mathbf{E}}^{(i)} \, | \bullet | \, \underline{\mathbf{A}}^{(i)} \in \mathbb{C}^{R_{i-1} R_{i-1}^E R_{i-1} \times R_i R_i^E R_i},$$

the left and right environment tensors are computed recursively by

$$\begin{aligned} \text{vec}\left(\mathcal{L}^{(<i)}\right)^{\text{H}} &= \text{vec}\left(\mathcal{L}^{(<i-1)}\right)^{\text{H}} \mathbf{Z}^{(i-1)} \in \mathbb{C}^{1 \times R_{i-1} R_{i-1}^E R_{i-1}}, \qquad i = 2, \cdots, n, \\ \text{vec}\left(\mathcal{R}^{(>i)}\right) &= \mathbf{Z}^{(i+1)} \, \text{vec}\left(\mathcal{R}^{(>i+1)}\right) \in \mathbb{C}^{R_i R_i^E R_i \times 1}, \qquad i = n, \cdots, N-1. \end{aligned} \tag{3.6}$$

Performing the contractions in the order shown in Figure 3.3, without explicitly forming $\mathbf{Z}^{(i)}$, each update of $\mathcal{L}^{(<i)}$ (or $\mathcal{R}^{(>i)}$) requires $\mathcal{O}\big(dR^3 R^E + d^2 R^2 (R^E)^2\big)$. Thus, evaluating the expectation value in Figure 3.2 costs $\mathcal{O}\big(\big(dR^3 R^E + d^2 R^2 (R^E)^2\big)(N+K)\big)$, assuming $R_n = R$ and $R_n^E = R^E$.

Single-site DMRG optimization. Assuming n-orthogonality, we can convert the global optimization problem in (3.4) into a set of linked small-scale optimization problems of the same form as

$$\begin{aligned} \min_{\underline{\mathbf{A}}^{(n)}} \quad & \underbrace{\frac{1}{2} \sum_{m=1}^{M} \left(y_m - \text{Tr}\left(\mathbf{A}_{[1,2,4;3]}^{(n)\text{H}} \, \mathbf{E}_{\text{eff},m}^{(n)} \, \mathbf{A}_{[1,2,4;3]}^{(n)} \right) \right)^2}_{=: f\left(\underline{\mathbf{A}}^{(n)}\right)}, \quad \text{s.t. } \|\underline{\mathbf{A}}^{(n)}\|_{\text{F}}^2 \leq 1, \\ \text{with} \quad & \mathbf{E}_{\text{eff},m}^{(n)} = \mathbf{A}_{\neq n}^{\text{H}} \, \mathbf{E}_{\text{TT},m} \, \mathbf{A}_{\neq n}, \quad \text{for } n = 1, 2, \cdots, N. \end{aligned} \tag{3.7}$$

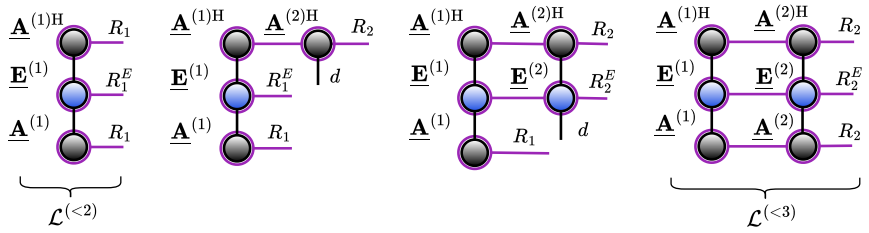

Fig. 3.3: Optimal ordering of the tensor network contractions needed to compute the left and right environments recursively; see (3.6).

Here, $\mathbf{A}^{(n)}_{[1,2,4;3]} \in \mathbb{C}^{R_{n-1}dR_n \times K}$ denotes the matrix unfolding of the nth core $\underline{\mathbf{A}}^{(n)}$. The unit-trace relaxation constraint reduces to a local constraint since $\|\mathbf{A}_{\mathrm{TT}}\|_{\mathrm{F}}^2 = \|\underline{\mathbf{A}}^{(n)}\|_{\mathrm{F}}^2$, thereby enabling efficient projection steps. The objective function above can be minimized, for example, using projected gradient descent (GD): $\underline{\mathbf{A}}^{(n)}_{i+1} = \Pi_{\mathcal{C}}\left(\underline{\mathbf{A}}^{(n)}_i - \eta \nabla f\left(\underline{\mathbf{A}}^{(n)}_i\right)\right)$, where $\Pi_{\mathcal{C}}$ denotes the projection onto the unit Frobenius ball $\mathcal{C} = \left\{\underline{\mathbf{X}} \;\middle|\; \|\underline{\mathbf{X}}\|_{\mathrm{F}}^2 \leq 1\right\}$, and

$$\nabla f(\underline{\mathbf{A}}^{(n)}) = -2 \sum_{m=1}^{M} \left(y_m - \mathrm{Tr}\left(\mathbf{A}^{(n)\mathrm{H}}_{[1,2,4;3]}\, \mathbf{E}^{(n)}_{\mathsf{eff},m}\, \mathbf{A}^{(n)}_{[1,2,4;3]}\right)\right) \mathbf{E}^{(n)}_{\mathsf{eff},m}\, \mathbf{A}^{(n)\mathrm{H}}_{[1,2,4;3]}$$

denotes the Euclidean gradient of the objective function f. We refer the reader to [4, 3, 45, 27, 19] for further details on solving such problems. More advanced methods, such as (projected) Gauss–Newton, can also be employed for constrained nonlinear least-squares problems [37]. After solving the local subproblem, the block index K is shifted to an adjacent core along with the orthogonality center, which becomes the next optimization target site. In this manner, all local sites are optimized sequentially across alternating sweeps until convergence. The overall procedure of the proposed method is outlined in Algorithm 3.1.

Rank adaptation and block-index sweeping in DMRG-I. During the left-to-right and right-to-left sweeps, the block index K is sequentially shifted to the active optimization block alongside the orthogonality center; this ensures that the active core always carries the block parameter K, which enables automatic (block)-rank adaptation. More specifically, assume a left-to-right sweep, where $\mathbf{A}_{\mathrm{TT}}$ is in Block–n–TT form and satisfies n–orthogonality. After optimizing the block-core $\underline{\mathbf{A}}^{(n)} \in \mathbb{C}^{R_{n-1}\cdot d\times K\cdot R_n}$, $\mathbf{A}_{\mathrm{TT}}$ is transformed to Block–$(n+1)$–TT form while also shifting the orthogonality center to $\underline{\mathbf{A}}^{(n+1)}$ using the following procedure (Line 6 of Algorithm 3.1). Let the δ-truncated SVD[5] of the unfolding of the nth core be $[\mathbf{U}, \mathbf{S}, \mathbf{V}] := \mathtt{SVD}_\delta\left(\mathbf{A}^{(n)}_{[1,2;3,4]}\right)$, where $\mathbf{U} \in \mathbb{C}^{R_{n-1}d\times R_n^{\mathrm{new}}}$. The TT-rank is updated as $R_n = R_n^{\mathrm{new}} = \mathrm{rank}\,(\mathbf{U})$, and $\mathbf{U}$ is reshaped into the new nth core with dimensions $R_{n-1}\cdot d \times R_n^{\mathrm{new}}$. The remaining factor $\mathbf{SV}$, which carries the index K, is absorbed into the next core and appropriately permuted to preserve the Block-TT core structure. Similarly, during the right-to-left sweep, K and the orthogonality center are moved to the preceding core.

Notice that for $K = 1$, the TT-rank cannot increase, since the updated rank is bounded by $R_n^{\mathrm{new}} \leq \min(R_{n-1}d, KR_n) = \min(R_{n-1}d, R_n)$. We show how this issue

[5] SVD with singular values below tolerance δ discarded.

Algorithm 3.1 Block-TT DMRG-I CS-QST

Input: Measurement data $\{y_m, \mathbf{E}_{\mathsf{TT},m}\}_{m=1}^M$, initial TT-rank $(R_0, \cdots, R_N)$, and K.
Output: $\mathbf{A}_{\mathsf{TT}}$ such that $\hat{\boldsymbol{\rho}}_{\mathsf{TT}} = \mathbf{A}_{\mathsf{TT}}\mathbf{A}_{\mathsf{TT}}^{\mathrm{H}}$.

1: Initialize $\mathbf{A}_{\mathsf{TT}}$ in Block–1–TT format `/* A_TT is in right-orthogonal form */`
2: Set the left and right boundary environments as $\{\mathcal{L}_m^{(<1)} = \mathcal{R}_m^{(>N)} = 1\}_{m=1}^M$.
3: Compute $\{\mathcal{R}_m^{(>N-1)}\}_{m=1}^M, \cdots, \{\mathcal{R}_m^{(>1)}\}_{m=1}^M$ recursively, as shown in (3.6).
repeat
`/* left-to-right half sweep */`
for $n = 1, 2, \cdots, N-1$ **do**
4: Compute $\{\mathbf{E}_{\mathsf{eff},m}^{(n)}\}_{m=1}^M$ via $\{\underline{\mathbf{E}}_m^{(n)}\}_{m=1}^M$, $\{\mathcal{L}_m^{(<n)}\}_{m=1}^M$ and $\{\mathcal{R}_m^{(>n)}\}_{m=1}^M$; see (3.5).
5: Minimize the local objective in (3.7) with respect to core $\underline{\mathbf{A}}^{(n)}$.
6: Transform Block–n–TT into Block–$(n+1)$–TT, so that the updated core is left-orthogonal and hence $\mathbf{A}_{\mathsf{TT}}$ is in $(n+1)$–orthogonal form, while updating the TT-rank.
`/* Current` $\underline{\mathbf{A}}^{(n+1)}$ `may serve as initialization for solving the` $(n+1)$`th core. */`
7: Apply projection: $\underline{\mathbf{A}}^{(n+1)} = \frac{\underline{\mathbf{A}}^{(n+1)}}{\max\left(1,\ \|\underline{\mathbf{A}}^{(n+1)}\|_F\right)}$. `/* ensures` $\|\mathbf{A}_{\mathsf{TT}}\|_F \leq 1$ `*/`
8: Compute $\{\mathcal{L}_m^{(<n+1)}\}_{m=1}^M$ via (3.6).
end for
`/* right-to-left half sweep */`
for $n = N, N-1, \cdots, 2$ **do**
9: Compute $\{\mathbf{E}_{\mathsf{eff},m}^{(n)}\}_{m=1}^M$ via $\{\underline{\mathbf{E}}_m^{(n)}\}_{m=1}^M$, $\{\mathcal{L}_m^{(<n)}\}_{m=1}^M$ and $\{\mathcal{R}_m^{(>n)}\}_{m=1}^M$; see (3.5).
10: Minimize the local objective in (3.7) with respect to core $\underline{\mathbf{A}}^{(n)}$.
11: Transform Block–n–TT into Block–$(n-1)$–TT, so that the updated core is right-orthogonal and hence $\mathbf{A}_{\mathsf{TT}}$ is in $(n-1)$–orthogonal form, while updating the TT-rank.
`/* Current` $\underline{\mathbf{A}}^{(n-1)}$ `may serve as initialization for solving the` $(n-1)$`th core. */`
12: Apply projection: $\underline{\mathbf{A}}^{(n-1)} = \frac{\underline{\mathbf{A}}^{(n-1)}}{\max\left(1,\ \|\underline{\mathbf{A}}^{(n-1)}\|_F\right)}$. `/* ensures` $\|\mathbf{A}_{\mathsf{TT}}\|_F \leq 1$ `*/`
13: Compute $\{\mathcal{R}_m^{(>n-1)}\}_{m=1}^M$ via (3.6).
end for
until a stopping criterion is met.
return $\mathbf{A}_{\mathsf{TT}}$

can be overcome within the two-site DMRG optimization scheme, following previous works in the literature [63, 52, 64, 23, 25, 10, 29].

Two-site DMRG optimization. The two-site DMRG-based QST method is a natural extension of the single-site DMRG-based QST method. In this approach, instead of optimizing one core at a time, two adjacent cores are combined into a larger core (a two-site core). This larger core is optimized and subsequently split back into two separate cores using truncated SVD. This method has been shown to achieve faster convergence per iteration [52, 25, 29] and allow automatic rank adaptation, even when $K = 1$. However, optimizing over a larger core tensor increases memory and computational requirements accordingly. The objective function can be expressed as:

$$\min_{\underline{\mathbf{A}}^{(n,n+1)}} \underbrace{\frac{1}{2}\sum_{m=1}^{M}\left(y_m - \mathrm{Tr}\left(\mathbf{A}_{[1,2,4;3]}^{(n,n+1)\mathrm{H}}\ \mathbf{E}_{\mathsf{eff},m}^{(n,n+1)}\ \mathbf{A}_{[1,2,4;3]}^{(n,n+1)}\right)\right)^2}_{=:f\left(\underline{\mathbf{A}}^{(n,n+1)}\right)}, \quad \text{s.t. } \|\underline{\mathbf{A}}^{(n,n+1)}\|_{\mathrm{F}}^2 \le 1, \tag{3.8}$$

$$\text{with} \quad \mathbf{E}_{\mathsf{eff},m}^{(n,n+1)} = \mathbf{A}_{\neq n,n+1}^{\mathrm{H}}\ \mathbf{E}_{\mathsf{TT},m}\ \mathbf{A}_{\neq n,n+1}, \quad \text{for } n = 1, 2, \cdots, N-1.$$

Here, $\mathbf{A}_{[1,2,4;3]}^{(n,n+1)} \in \mathbb{C}^{R_{n-1}d^2R_{n+1}\times K}$ denotes the matrix unfolding of the two-site core $\underline{\mathbf{A}}^{(n,n+1)} = \underline{\mathbf{A}}^{(n)} \,|\!\otimes\!|\, \underline{\mathbf{A}}^{(n+1)}$. The two-site effective operator is computed by contracting the tensorized two-site local operator $\mathcal{E}^{(n,n+1)} \in \mathbb{C}^{R_{n-1}^E\times d^2\times d^2\times R_{n+1}^E}$, obtained from the strong Kronecker product of the corresponding cores $\underline{\mathbf{E}}^{(n)} \,|\!\otimes\!|\, \underline{\mathbf{E}}^{(n+1)}$, with the left and right environments, as follows:

$$\mathcal{Z} = \left(\mathcal{E}^{(n,n+1)} \bullet_1^2 \mathcal{L}^{(<n)}\right) \bullet_3^2 \mathcal{R}^{(>n+1)},\ \mathbf{E}_{\mathsf{eff}}^{(n,n+1)} = \mathbf{Z}_{[3,1,5;4,2,6]} \in \mathbb{C}^{R_{n-1}d^2R_{n+1}\times R_{n-1}d^2R_{n+1}}. \tag{3.9}$$

Rank adaptation and block-index sweeping in DMRG-II. Similarly, during the left-to-right and right-to-left sweeps, the index K is shifted between successive two-site cores so that it always remains attached to the active optimization block. Assuming a left-to-right sweep, after optimizing the two-site core $\underline{\mathbf{A}}^{(n,n+1)}$, the index K is shifted to the next position by splitting the supercore back into two cores via a truncated SVD. This simultaneously converts a Block–n–TT into a Block–$(n+1)$–TT and moves the orthogonality center to the $(n+1)$th core. Let $\mathcal{Z}^{(n,n+1)} \in \mathbb{C}^{R_{n-1}\times d\times d\times K\times R_{n+1}}$ denote the tensor corresponding to $\underline{\mathbf{A}}^{(n,n+1)}$, and consider the δ-truncated SVD of its matrix unfolding: $[\mathbf{U},\mathbf{S},\mathbf{V}] := \mathtt{SVD}_\delta\left(\mathbf{Z}_{[1,2;3,4,5]}^{(n,n+1)}\right)$, where $\mathbf{U} \in \mathbb{C}^{R_{n-1}d\times R_n^{\mathrm{new}}}$. The TT-rank is updated as $R_n = R_n^{\mathrm{new}} = \mathrm{rank}\,(\mathbf{U})$. The factor $\mathbf{U}$ is then reshaped into the new nth core of dimensions $R_{n-1}\cdot d \times R_n^{\mathrm{new}}$, while $\mathbf{SV}$ is reshaped into the new $(n+1)$th core of dimensions $R_n^{\mathrm{new}}\cdot d\times K\cdot R_{n+1}$. This transfers the index K and shifts the orthogonality center to the $(n+1)$th core.

Notice that when $K = 1$ the two-site DMRG allows rank adaptation since the updated rank is now bounded by $R_n^{\mathrm{new}} \le \min(R_{n-1}d, dKR_{n+1}) = \min(R_{n-1}d, dR_{n+1})$. In contrast to single-site DMRG, the two-site DMRG scheme allows for a larger increase in the TT-rank when needed. However, this added flexibility comes at the cost of larger local optimization problems, which may slow down the overall optimization per iteration. The overall procedure is summarized in Algorithm 3.2.

Overlapping measurement operators for efficient environment caching. Consider a family of measurement operators of the form,

$$\mathbf{E}_{\mathsf{TT}} = \underbrace{\underline{\mathbf{E}}^{(1)} \,|\!\otimes\!|\, \cdots \,|\!\otimes\!|\, \underline{\mathbf{E}}^{(n_1-1)}}_{\text{fixed left}} \,|\!\otimes\!|\, \underbrace{\underline{\mathbf{E}}^{(n_1)} \,|\!\otimes\!|\, \cdots \,|\!\otimes\!|\, \underline{\mathbf{E}}^{(n_2)}}_{\text{variable block}} \,|\!\otimes\!|\, \underbrace{\underline{\mathbf{E}}^{(n_2+1)} \,|\!\otimes\!|\, \cdots \,|\!\otimes\!|\, \underline{\mathbf{E}}^{(N)}}_{\text{fixed right}}$$

where the left cores $\{\underline{\mathbf{E}}^{(i)}\}_{i<n_1}$ and the right cores $\{\underline{\mathbf{E}}^{(i)}\}_{i>n_2}$ are the same for the whole family, while the operator cores inside the block $[n_1, n_1+1, \cdots, n_2]$ vary (the single-qudit case corresponds to $n_1 = n_2$). Here, we pre-contract the fixed left and right parts into environment tensors $\mathcal{L}^{(<n_1)}$ and $\mathcal{R}^{(>n_2)}$. The expectation value of any member in this family then reduces to a contraction over only the variable block tensors. This enables efficient environment caching and expectation value evaluation.

To perform QST with this scheme, one can optimize the cores of $\mathbf{A}_{\mathsf{TT}}$ over the measurements, with the variable block shifted left or right in a sliding-window fashion. Only the cores corresponding to the variable block can be optimized. This allows local

Algorithm 3.2 Block-TT DMRG-II CS-QST

Input: Measurement data $\{y_m, \mathbf{E}_{\mathrm{TT},m}\}_{m=1}^{M}$, initial TT-rank $(R_0, \cdots, R_N)$, and K.
Output: $\mathbf{A}_{\mathrm{TT}}$ such that $\hat{\boldsymbol{\rho}}_{\mathrm{TT}} = \mathbf{A}_{\mathrm{TT}}\mathbf{A}_{\mathrm{TT}}^{\mathrm{H}}$.

1: Initialize $\mathbf{A}_{\mathrm{TT}}$ in Block–1–TT format `/* A_TT is in right-orthogonal form */`
2: Set the left and right boundary environments as $\{\mathcal{L}_m^{(<1)} = \mathcal{R}_m^{(>N)} = 1\}_{m=1}^{M}$.
3: Compute $\{\mathcal{R}_m^{(>N-1)}\}_{m=1}^{M}, \cdots, \{\mathcal{R}_m^{(>1)}\}_{m=1}^{M}$ recursively, as shown in (3.6).
repeat
  `/* left-to-right half sweep */`
  for $n = 1, 2, \cdots, N-1$ **do**
    4: Compute $\{\mathbf{E}_{\mathsf{eff},m}^{(n,n+1)}\}_{m=1}^{M}$ via $\{\underline{\mathbf{E}}_m^{(n)} \boxtimes \underline{\mathbf{E}}_m^{(n+1)}\}_{m=1}^{M}$, $\{\mathcal{L}_m^{(<n)}\}_{m=1}^{M}$ and $\{\mathcal{R}_m^{(>n+1)}\}_{m=1}^{M}$, as shown in (3.9).
    5: Minimize the local objective in (3.8) with respect to the two-site core $\underline{\mathbf{A}}^{(n)}$.
    6: Split the updated two-site core $\underline{\mathbf{A}}^{(n,n+1)}$ via a truncated SVD, updating the TT-rank, shifting the orthogonality center, and converting the Block–n–TT into a Block–$(n+1)$–TT.
    7: Apply projection: $\underline{\mathbf{A}}^{(n+1)} = \frac{\underline{\mathbf{A}}^{(n+1)}}{\max\left(1,\ \|\underline{\mathbf{A}}^{(n+1)}\|_F\right)}$. `/* ensures ||A_TT||_F ≤ 1 */`
    8: Compute $\{\mathcal{L}_m^{(<n+1)}\}_{m=1}^{M}$ via (3.6).
  end for
  `/* right-to-left half sweep */`
  for $n = N, N-1, \cdots, 2$ **do**
    Carry out the right-to-left half-sweep similarly.
  end for
until a stopping criterion is met.
return $\mathbf{A}_{\mathrm{TT}}$

POVMs to be measured on a subset of qubits while the remaining qubits are kept in a fixed basis. The approach is particularly useful for one-dimensional systems in which correlations are primarily restricted to neighbouring qubits. Similar ideas have been proposed, including efficient MPS tomography protocols for many-body systems based on local reductions (i.e., partial traces of the density matrix) [38, 28, 17, 56].

Furthermore, if the left and right cores (the fixed parts) of the measurement operators are identity matrices, i.e., $\{\underline{\mathbf{E}}^{(i)} = [\mathbf{I}_d] \in \mathbb{C}^{d\times d}\}_{i<n_1 \text{ or } i>n_2}$, evaluating the full-state trace reduces to evaluating the trace on the reduced density matrix after tracing out fixed subsystems, as shown in Figure 3.4.

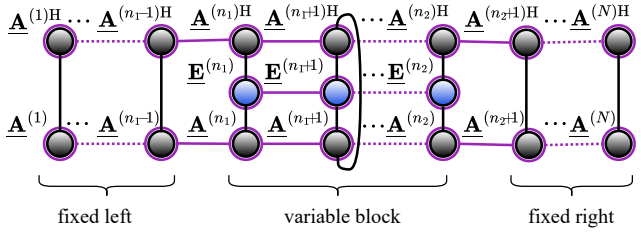

Fig. 3.4: Expectation value evaluation on a local subsystem, equal to that obtained from the reduced density matrix after tracing out the fixed parts.

Note that our approach differs fundamentally from the MPS tomography methods

proposed in [38, 28, 17], which aim to reconstruct the full density matrix from exclusively local reduced density matrices. In contrast, our method estimates expectation values from (full) trace-based measurements and leverages a Block-TT parameterization that is directly applicable to mixed states.

Implementation details. Algorithms 3.1 and 3.2 are implemented in the PyTorch framework, using an L-BFGS algorithm to solve the local optimization problems (3.7) and (3.8). A link to the repository will be made available upon publication.

3.3. Computational complexity. To assess the computational complexity of Algorithms 3.1 and 3.2, we consider an N-qudit system and assume uniform ranks for the Block-TT tensor $\mathbf{A}_{\mathrm{TT}}$ and the measurement operators $\mathbf{E}_{\mathrm{TT}}$, i.e., $R_n = R$ and $R_n^E = R^E$. We report the complexity per sweep, assuming that the ranks and block index K remain bounded during optimization. The dominant computational cost arises from (i) constructing the effective operators for the M measurements (Line 4) and (ii) solving the reduced optimization problem (Line 5), assuming N_{eval} evaluations of the objective function and its gradient are required to satisfy the stopping criterion. The corresponding complexities are summarized in Table 3.1. The costs associated with the orthogonalization and truncation steps (Lines 6 and 7) are negligible, as these operations act only on one-site or two-site cores and are independent of M. Moreover, owing to the efficient contraction strategy described in Subsection 3.2, the update of the left and right environments (Line 8) does not constitute the dominant computational cost.

Table 3.1: Computational complexity per sweep (flop) of Algorithms 3.1 and 3.2, where N_{eval} denotes the number of objective and gradient evaluations required by the chosen local solver.

Dominant Terms	**DMRG-I**	**DMRG-II**
Effective operators (line 4)	$\mathcal{O}\big(NM\left(d^2R^4R^E + d^2R^2(R^E)^2\right)\big)$	$\mathcal{O}\big(NM\left(d^4R^4R^E + d^4R^2(R^E)^2\right)\big)$
Solving local problem (line 5)	$\mathcal{O}\big(NN_{\mathrm{eval}}Md^2R^4K\big)$	$\mathcal{O}\big(NN_{\mathrm{eval}}Md^4R^4K\big)$

Per measurement, the construction of the effective operator at a given site (DMRG-I) or site pair (DMRG-II) requires $\mathcal{O}\left(d^2R^4R^E + d^2R^2(R^E)^2\right)$ operations for DMRG-I (Equation (3.5)) and $\mathcal{O}\left(d^4R^4R^E + d^4R^2(R^E)^2\right)$ for DMRG-II (Equation (3.9)). Furthermore, each evaluation of the local objective function and its gradient incurs a cost of $\mathcal{O}\left(Md^2R^4K\right)$ for DMRG-I and $\mathcal{O}\left(Md^4R^4K\right)$ for DMRG-II.

4. Numerical experiments. This section evaluates the performance of the proposed methods across the following three experimental settings:

- First, we show the accuracy and computational efficiency of the proposed approaches by comparing them with standard low-rank baselines: the rank-minimization formulation (2.1), solved using the splitting conic solver in CVXPY [9], and the error-minimization approach (2.2) based on the Burer–Monteiro factorization [4], as implemented in [27]. Performance is assessed using fidelity, $\left(\mathrm{Tr}\left(\sqrt{\sqrt{\boldsymbol{\rho}}\hat{\boldsymbol{\rho}}\sqrt{\boldsymbol{\rho}}}\right)\right)^2$, and trace distance, $\frac{1}{2}\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|_1$, which quantify (overlap) similarity and distinguishability, respectively.
- Second, we compare the two DMRG methods in terms of rank adaptivity, computational time, and reconstruction accuracy as the number of qubits increases. Here, performance is primarily evaluated using the Frobenius norm

between the ground-truth and reconstructed states, as well as the accuracy of expectation values computed for unseen (test) measurement operators.
- Finally, we showcase that overlapping operators enable large-scale QST for structured states whose correlations are predominantly local, i.e., restricted to neighbouring qubits. As an illustrative example, we consider a structured MPS state and evaluate reconstruction accuracy in terms of fidelity.

The experiments were performed on an HP EliteBook 845 G8 Notebook PC equipped with an AMD Ryzen 7 PRO 5850U processor and 32 GB of RAM.

4.1. Accuracy and computational efficiency. A random complex mixed state of an N-qubit system is generated by sampling a Block-TT tensor $\mathbf{A}_{\mathrm{TT}}$ with $K = 2$ and a TT-rank with maximum entry 3. The entries of each core tensor are complex numbers whose real and imaginary components are drawn independently from a standard normal distribution. After normalizing $\mathbf{A}_{\mathrm{TT}}$ to unit norm, the corresponding TTM density operator is constructed as $\boldsymbol{\rho}_{\mathrm{TT}} = \mathbf{A}_{\mathrm{TT}}\mathbf{A}_{\mathrm{TT}}^{\mathrm{H}}$.

The number of qubits N is varied from 4 to 7. For each system size, local SIC-POVM measurements are performed, where each N-qubit measurement operator is formed by taking the Kronecker product of N single-qubit SIC-POVM elements. The total number of measurement operators scales as $M = \alpha \cdot P \log(N)$, where $P = \sum_{n=1}^{N} R_{n-1}^2 d^2 R_n^2$ denotes the parameters of the equivalent MPO representation [48, 49]. This logarithmic scaling is directly motivated by the logarithm of the covering number (ϵ-entropy), which quantifies the metric complexity of the underlying state class [48, 49] and provides a useful basis for designing QST evaluation experiments. The sampling factor α is varied from 0.25 to 1.0, and Gaussian noise is added to the measurements with signal-to-noise ratio (SNR) equal to 60 dB. The proposed algorithms (Algorithms 3.1 and 3.2) are executed alongside reference methods, including a convex optimization approach (CVX) [16, 9] and a low-rank Burer–Monteiro solver (BM) [27]. For the DMRG algorithms, the relative objective function tolerance is set to 10^{-4}, chosen below the expected 10^{-3} relative noise level corresponding to an SNR of 60 dB, with a maximum of 5 sweeps. Reference methods use the same tolerance, terminating after a maximum of 400 iterations for CVX and 2000 for BM if convergence is not reached. The median fidelity and trace distance across 10 trials are reported in Figure 4.1.

The proposed DMRG algorithms consistently outperform CVX and BM in both accuracy and efficiency. By exploiting the TTM structure, DMRG-I and DMRG-II achieve near-perfect fidelity (≥ 0.99 for $\alpha \geq 0.5$) and significantly lower trace distance error, while running over an order of magnitude faster. Across all methods, increasing the sampling factor α enhances reconstruction accuracy with minor computational cost. Notably, at low sampling rates ($\alpha = 0.25$), DMRG-I slightly outperforms DMRG-II because limited data allows smaller single-core subproblems to be solved more accurately than larger two-core ones; as α increases, DMRG-II achieves superior performance. Finally, performance degradation at larger N is primarily due to fixed sweep and iteration budgets, which can be resolved by scaling maximum iterations with system size.

4.2. Rank adaptivity and scalability. In the first experiment, the rank adaptivity and convergence of the proposed DMRG-I and DMRG-II algorithms are evaluated. A random $N = 7$-qubit ground truth state is generated in Block-TT form with TT-rank $(1, 2, 3, 9, 10, 4, 2, 1)$ and $K = 2$, and local SIC-POVM measurements are performed with $M = 0.5 \cdot P \log(N)$ measurements. Gaussian noise is added to the measurements with an SNR = 60 dB. QST is then performed using both methods,

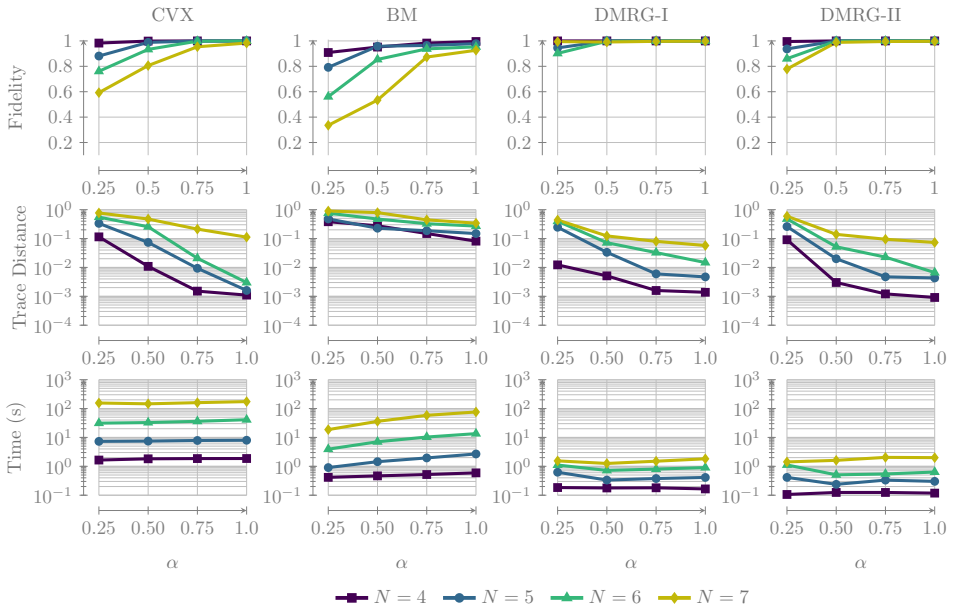

Fig. 4.1: Baseline comparison of reconstruction fidelity, trace distance, and runtime versus sampling rate α for system sizes $N \in \{4, 5, 6, 7\}$. The proposed DMRG algorithms consistently outperform standard baselines, achieving near-perfect fidelity (≥ 0.99 for $\alpha \geq 0.5$) and over $10\times$ speedup at $N = 7$.

recording the objective loss per site in DMRG-I and per pair of consecutive sites in DMRG-II. The maximum number of sweeps is set to 3 in both methods, and the relative objective function tolerance is set to 10^{-4} (according to the noise level). The same random initialization is used in both methods, with the TT-rank initialized to $(1, 1, \ldots, 1)$. During optimization, the maximum element of the TT-rank, $R_{\max}$, is recorded. Median results are computed over 10 independent trials and are plotted in Figure 4.2. It is observed that DMRG-II achieves faster convergence within one full sweep, whereas DMRG-I achieves similar convergence in two sweeps. As shown in the inset plot of Figure 4.2, ranks adapt faster in DMRG-II, whereas rank growth in DMRG-I is $\mathcal{O}(K)$. Thus, for small K, the rank adapts slowly, and more sweeps are required to converge.

Next, the number of qubits N is varied from 4 to 12. As in previous experiments, the ground-truth state is generated, local SIC-POVM measurements are performed, and outcomes are corrupted with an SNR = 60 dB. QST is executed with a relative objective tolerance of 10^{-4} and up to 10 sweeps. Three metrics are evaluated: relative Frobenius reconstruction error, relative norm distance between true and recovered expectation values on a test set ($\frac{1}{4}$ of the training set size), and computation time to satisfy the stopping criterion. Median results over 10 independent trials are plotted in Figure 4.3. While higher accuracy is achieved by DMRG-II for smaller systems, the accuracy gap between DMRG-I and DMRG-II vanishes as system size increases. A similar crossover is observed in computation time: DMRG-II converges slightly faster for smaller systems, but becomes slower as N grows. This slowdown occurs because faster TT-rank growth in DMRG-II increases the cost of solving two-site cores and

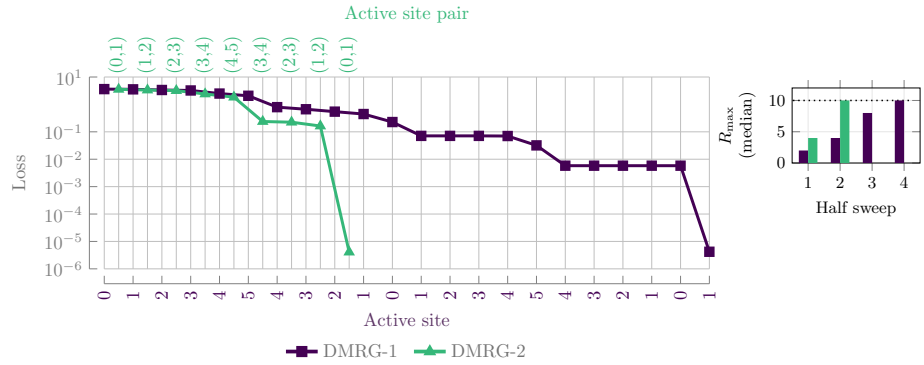

Fig. 4.2: Convergence and rank adaptivity comparison between DMRG-I and DMRG-II ($N = 7$, true $R_{\max} = 10$). The main plot shows the median loss across visited sites/site-pairs, with DMRG-II converging within a single sweep. The inset shows that DMRG-II reaches the target maximum rank ($R_{\max} = 10$, dotted line) by sweep 1, whereas DMRG-I adapts more gradually over two sweeps.

performing SVDs. Furthermore, a general increase in reconstruction error is observed for larger systems due to the fixed sweep count; maintaining convergence requires both iterations and measurement counts to scale strongly with N.

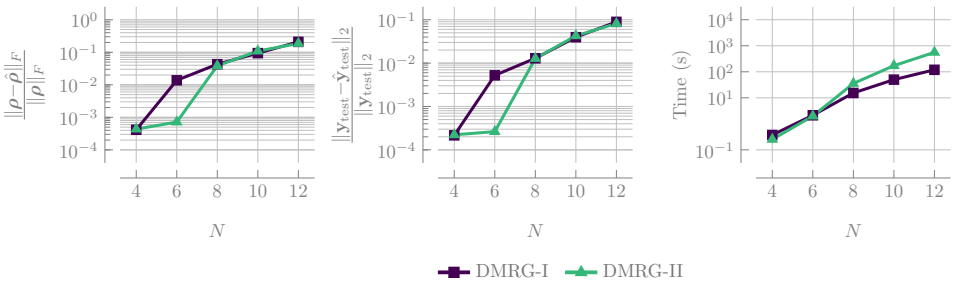

Fig. 4.3: Scalability comparison between DMRG-I and DMRG-II across qubit counts N: Frobenius error (left), test prediction error (middle), and computation time (right). While DMRG-II achieves higher accuracy and speed for small systems ($N \leq 6$), its accuracy advantage vanishes for larger N, where its runtime increases significantly faster than DMRG-I.

Finally, the impact of the block parameter K on estimation accuracy and runtime is evaluated to demonstrate performance trade-offs when tuning model capacity. Random N-qubit ground-truth states ($N \in \{4, 5, 6, 7, 8\}$) are generated in Block-TT form with TT-ranks $(1, 3, \ldots, 3, 1)$ and fixed true capacity $K_{\text{true}} = 2$. Local SIC-POVM measurements are performed with $M = 0.2 \cdot P \log(N)$, and the measurement outcomes are corrupted with additive noise corresponding to an SNR = 60 dB. DMRG-II is then used to reconstruct the state for $K \in \{1, 2, 3, 4\}$. The maximum sweep count is set to 4 with a relative objective tolerance of 10^{-5}. Median fidelity and computation time over 20 independent trials are plotted in Figure 4.4. As K increases from 1 to-

ward K_{true}, state fidelity increases significantly, achieving near-unit fidelity across all system sizes. Smaller K values require slightly longer optimization times; however, larger values converge faster. Notably, at $K = 2$, reconstruction achieves the best fidelity, whereas when K exceeds the true value, fidelity decreases slightly because extra parameters tend to fit the noise.

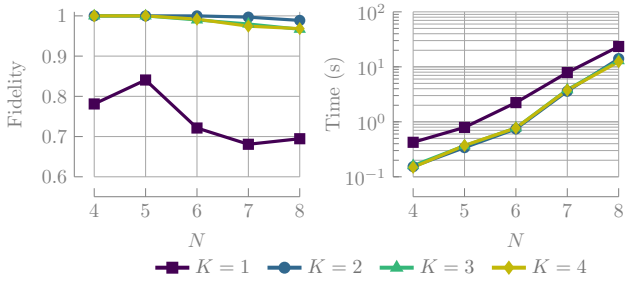

Fig. 4.4: Reconstruction fidelity (left) and computation time (right) versus qubit count N across model capacities K. Accuracy peaks when $K = K_{\text{true}}$, and runtime increases with system size N.

4.3. Exploiting overlapping measurements for large-scale QST. To show that overlapping localized measurements enable large-scale QST, we prepare a target state using the single-site DMRG algorithm implemented in PyTeNet [33]. Specifically, we obtain the ground state of an $N = 30$ site 1D transverse-field Ising model Hamiltonian:

$$\mathbf{H} = -J \sum_{n=1}^{N-1} \boldsymbol{\sigma}_z^{(n)} \boldsymbol{\sigma}_z^{(n+1)} - g \sum_{n=1}^{N} \boldsymbol{\sigma}_x^{(n)}, \tag{4.1}$$

with parameters $J = 1.0$ and $g = 2.0$, where $\boldsymbol{\sigma}_\alpha^{(n)} = \mathbf{I}_2^{\otimes n-1} \otimes \boldsymbol{\sigma}_\alpha \otimes \mathbf{I}_2^{\otimes N-n}$ is the $2^N \times 2^N$ operator obtained by placing the single-site Pauli matrix $\boldsymbol{\sigma}_\alpha$ at site n, with $\alpha \in \{z, x\}$. The target state is obtained as an MPS/TT with the maximum element of the TT-rank $R_{\max} = 6$.

We construct rank-1 TT measurement operators $\mathbf{E}_{\text{TT}}$ localized to a spatial window $[n_1, n_2]$:

$$\mathbf{E}_{\text{TT}} = \underbrace{\mathbf{I}_2^{\otimes n_1-1}}_{\text{fixed}} \otimes \underbrace{\Big(\mathbf{E}(\mathbf{n}_{n_1}) \otimes \cdots \otimes \mathbf{E}(\mathbf{n}_{n_2})\Big)}_{\text{variable block}} \otimes \underbrace{\mathbf{I}_2^{\otimes N-n_2}}_{\text{fixed}},$$

where $\mathbf{E}(\mathbf{n}_i) = \frac{1}{2}\left(\mathbf{I}_2 + n_{ix}\boldsymbol{\sigma}_x + n_{iy}\boldsymbol{\sigma}_y + n_{iz}\boldsymbol{\sigma}_z\right)$, with $\mathbf{n}_i = (n_{ix}, n_{iy}, n_{iz})^\top$ uniformly distributed over the surface of the Bloch sphere (solid angle 4π) via $n_{ix} = \sin\alpha_i \cos\varphi_i$, $n_{iy} = \sin\alpha_i \sin\varphi_i$, $n_{iz} = \cos\alpha_i$, with $\cos\alpha_i \sim \text{Uniform}[-1, 1]$, $\varphi_i \sim \text{Uniform}[0, 2\pi)$. Notice that the pair $\{\mathbf{E}(\mathbf{n}_i), \mathbf{E}(-\mathbf{n}_i)\}$ forms a valid two-outcome projective POVM satisfying $\mathbf{E}(\mathbf{n}_i) + \mathbf{E}(-\mathbf{n}_i) = \mathbf{I}_2$.

Measurement operators are generated by sweeping a 1D window of width $w = n_2 - n_1 + 1$ across the N-qubit chain with stride s. At each window position $[n_1, n_2]$, a batch of product measurement operators is sampled: the local Bloch vectors $\{\mathbf{n}_i\}_{i=n_1}^{n_2}$ are drawn independently and uniformly at random, while every site outside the window is assigned the identity $\mathbf{I}_2$. Shifting the window by s sites gives a spatial overlap

of $o = w - s$ sites between consecutive windows ($s \leq w$), or equivalently a relative overlap of $(1 - s/w) \times 100\%$. We study four window–stride configurations, $(w, s) \in \{(3,3), (6,3), (4,2), (4,1)\}$, with overlap percentages of 0%, 50%, 50%, and 75%, respectively. For accurate reconstruction, the window width and overlap must be large enough to jointly capture all site pairs that are significantly correlated in the target ground state.

In the first experiment, the total number of measurement operators is fixed at 2400 for every configuration. The number of window positions is $N_p = \left\lfloor \frac{N-w}{s} \right\rfloor + 1$, so each position receives approximately $\lfloor 2400/N_p \rfloor$ operators. After generation, the expectation values are corrupted by additive white noise at SNR = 60 dB. The DMRG-I algorithm then reconstructs the quantum state. Left and right environment tensors for the identity operators outside the active window are computed once and cached; since consecutive window positions share identical operators on all sites outside their union, the corresponding environment tensors are reused without recomputation. Table 4.1 shows that, under the same total measurement budget, the choice of (w, s) leads to substantially different reconstruction fidelity and computation time.

Table 4.1: Reconstruction fidelity and execution time (mean ± std over 10 trials) for a fixed total budget of 2400 measurement operators. Zero overlap yields substantially lower fidelity, while sufficient overlap enables accurate reconstruction

s	w	Overlap	N_p	Fidelity	Time (s)
3	3	0%	10	0.7026 ± 0.0165	146.04 ± 31.23
3	6	50%	9	0.9616 ± 0.0171	227.99 ± 42.22
2	4	50%	14	0.9750 ± 0.0032	208.04 ± 14.59
1	4	75%	27	0.9840 ± 0.0037	221.95 ± 8.26

In the second experiment, each window position is assigned exactly 100 measurement operators, so the total count $100N_p$ varies across configurations. Under the same four configurations, we compare reconstruction fidelity and computation time. Table 4.2 summarizes the results.

Table 4.2: Reconstruction fidelity and execution time (mean ± std over 10 trials) for a fixed 100 operators per window position. By tuning window width and stride, one can trade off between reconstruction fidelity and computational cost, provided the total measurement budget remains sufficient.

s	w	Overlap	N_p	Fidelity	Time (s)
3	3	0%	10	0.6941 ± 0.0174	64.95 ± 6.31
3	6	50%	9	0.7990 ± 0.0349	125.77 ± 12.92
2	4	50%	14	0.9697 ± 0.0038	110.09 ± 7.03
1	4	75%	27	0.9854 ± 0.0025	196.21 ± 3.91

By tuning the window width and stride, one can trade off reconstruction fidelity against computational cost; however, sufficient spatial overlap and total measurement budget must both be maintained to ensure that all physically relevant correlations of the target state are adequately captured.

5. Conclusion. We introduced a Block-TT Burer–Monteiro framework for quantum state tomography of high-dimensional mixed quantum states. The proposed parameterization generalizes the Burer–Monteiro factorization to high-dimensional density matrices with a parameter count scaling linearly in the number of qudits, making it both memory- and computation-efficient. Building on this framework, we developed rank-adaptive single-site and two-site DMRG algorithms for learning low-rank mixed states. The resulting approach produces physically valid states by construction; thereby avoiding explicit positivity constraints. We further derived efficient tensor-network contraction schemes for local core optimization and introduced overlapping measurement operators together with effective environment caching strategies to enable scalable tomography of large structured quantum systems. Numerical experiments demonstrate that the proposed framework can accurately reconstruct low-rank quantum states from limited measurements while offering substantial reductions in storage and computational requirements.

REFERENCES

[1] S. Aaronson, *Shadow tomography of quantum states*, in Proceedings of the 50th annual ACM SIGACT symposium on theory of computing, 2018, pp. 325–338.

[2] C. H. Baldwin, I. H. Deutsch, and A. Kalev, *Strictly-complete measurements for bounded-rank quantum-state tomography*, Phys. Rev. A, 93 (2016), p. 052105.

[3] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, *Dropping convexity for faster semi-definite optimization*, in Conference on Learning Theory, PMLR, 2016, pp. 530–582.

[4] S. Burer and R. D. C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Math. Program., 95 (2003), pp. 329–357.

[5] L. Calderaro, G. Foletto, D. Dequal, P. Villoresi, and G. Vallone, *Direct reconstruction of the quantum density matrix by strong measurements*, Phys. Rev. Lett., 121 (2018), p. 230501.

[6] A. Cichocki, N. Lee, I. Oseledets, A.-H. Phan, Q. Zhao, D. P. Mandic, et al., *Tensor networks for dimensionality reduction and large-scale optimization: Part 1: low-rank tensor decompositions*, Found. Trends Mach. Learn., 9 (2016), pp. 249–429.

[7] M. Cramer, M. B. Plenio, S. T. Flammia, R. Somma, D. Gross, S. D. Bartlett, O. Landon-Cardinal, D. Poulin, and Y.-K. Liu, *Efficient quantum state tomography*, Nat. Commun., 1 (2010), p. 149.

[8] G. D. l. Cuevas, N. Schuch, D. Pérez-García, and J. Ignacio Cirac, *Purifications of multipartite states: limitations and constructive methods*, New J. Phys., 15 (2013), p. 123021.

[9] S. Diamond and S. Boyd, *Cvxpy: A python-embedded modeling language for convex optimization*, J. Mach. Learn. Res., 17 (2016).

[10] S. V. Dolgov, B. N. Khoromskij, I. Oseledets, and D. V. Savostyanov, *Computation of extreme eigenvalues in higher dimensions using block tensor train format*, Comput. Phys. Commun., 185 (2014), pp. 1207–1216.

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

[12] M. Fazel, *Matrix rank minimization with applications*, PhD thesis, Stanford University, 2002.

[13] M. Fazel, H. Hindi, and S. Boyd, *Rank minimization and applications in system theory*, in Proceedings of the 2004 ACC, vol. 4, IEEE, 2004, pp. 3273–3278.

[14] T. Feng, C. Ren, and X. Zhou, *Direct measurement of density-matrix elements using a phase-shifting technique*, Phys. Rev. A, 104 (2021), p. 042403.

[15] L. Grasedyck, M. Kluge, and S. Krämer, *Variants of alternating least squares tensor completion in the tensor train format*, SIAM J. Sci. Comput., 37 (2015), pp. A2424–A2450.

[16] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing*, Phys. Rev. Lett., 105 (2010), p. 150401.

[17] Y. Guo and S. Yang, *Quantum state tomography with locally purified density operators and local measurements*, Commun. Phys., 7 (2024), p. 322.

[18] S. Holtz, T. Rohwedder, and R. Schneider, *The alternating linear scheme for tensor optimization in the tensor train format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713.

[19] E.-L. Hu and J. T. Kwok, *Low-rank matrix learning using biconvex surrogate minimization*, IEEE Transactions on Neural Networks and Learning Systems, 30 (2019), pp. 3517–3527.

[20] H.-Y. Huang, R. Kueng, and J. Preskill, *Predicting many properties of a quantum system from very few measurements*, Nat. Phys., 16 (2020), pp. 1050–1057.

[21] A. Kalev, R. L. Kosut, and I. H. Deutsch, *Quantum tomography protocols with positivity are compressed sensing protocols*, npj Quantum Inf., 1 (2015), pp. 1–6.

[22] V. Khoromskaia and B. N. Khoromskij, *Tensor numerical methods in quantum chemistry*, Walter de Gruyter GmbH & Co KG, 2018.

[23] B. N. Khoromskij and I. Oseledets, *DMRG+QTT approach to computation of the ground state for the molecular schrödinger operator*, MPI MIS Preprint 69/2010, Max Planck Institute for Mathematics in the Sciences (MPI MIS), 2010.

[24] J. L. Kim, G. Kollias, A. Kalev, K. X. Wei, and A. Kyrillidis, *Fast quantum state reconstruction via accelerated non-convex programming*, Photonics, 10 (2023).

[25] D. Kressner, M. Steinlechner, and A. Uschmajew, *Low-rank tensor methods with subspace correction for symmetric eigenvalue problems*, SIAM J. Sci. Comput., 36 (2014), pp. A2346–A2368.

[26] S. Kuzmin, V. Mikhailova, I. Dyakonov, and S. Straupe, *Learning the tensor network model of a quantum state using a few single-qubit measurements*, Phys. Rev. A, 109 (2024), p. 052616.

[27] A. Kyrillidis, A. Kalev, D. Park, S. Bhojanapalli, C. Caramanis, and S. Sanghavi, *Provable compressed sensing quantum state tomography via non-convex methods*, npj Quantum Inf., 4 (2018), p. 36.

[28] B. P. Lanyon, C. Maier, M. Holzäpfel, T. Baumgratz, C. Hempel, P. Jurcevic, I. Dhand, A. Buyskikh, A. J. Daley, M. Cramer, et al., *Efficient tomography of a quantum many-body system*, Nat. Phys., 13 (2017), pp. 1158–1162.

[29] N. Lee and A. Cichocki, *Estimating a few extreme singular values and vectors for large-scale matrices in tensor train format*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 994–1014.

[30] N. Lee and A. Cichocki, *Fundamental tensor operations for large-scale data analysis using tensor network formats*, Multidimens. Syst. Signal Process., 29 (2018), pp. 921–960.

[31] W.-T. Liu, T. Zhang, J.-Y. Liu, P.-X. Chen, and J.-M. Yuan, *Experimental quantum state tomography via compressed sampling*, Phys. Rev. Lett., 108 (2012), p. 170403.

[32] Y.-K. Liu, *Universal low-rank matrix recovery from pauli measurements*, Adv. Neural Inf. Process. Syst., 24 (2011).

[33] C. B. Mendl, *Pytenet: A concise python implementation of quantum tensor network algorithms*, J. Open Source Softw., 3 (2018), p. 948.

[34] T. Moroder, P. Hyllus, G. Tóth, C. Schwemmer, A. Niggebaum, S. Gaile, O. Gühne, and H. Weinfurter, *Permutationally invariant state reconstruction*, New J. Phys., 14 (2012), p. 105001.

[35] J. Morris and B. Dakić, *Selective quantum state tomography*, arXiv preprint arXiv:1909.05880, (2019).

[36] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, Cambridge university press, 2010.

[37] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 2006.

[38] R. O'Donnell and J. Wright, *Efficient quantum tomography*, in Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing, 2016, pp. 899–912.

[39] R. Orús, *A practical introduction to tensor networks: Matrix product states and projected entangled pair states*, Ann. Phys., 349 (2014), pp. 117–158.

[40] I. Oseledets, *Approximation of $2^d \times 2^d$ matrices using tensor decomposition*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2130–2145.

[41] I. Oseledets, *Dmrg approach to fast linear algebra in the tt-format*, Comput. Meth. Appl. Mat., 11 (2011), pp. 382–393.

[42] I. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.

[43] M. Paris, *The modern tools of quantum mechanics: A tutorial on quantum states, measurements, and operations*, Eur. Phys. J. Spec. Top., 203 (2012), pp. 61–86.

[44] M. Paris and J. Rehacek, *Quantum State Estimation*, Lecture Notes in Physics, Springer Berlin Heidelberg, 2004.

[45] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, *Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably*, SIAM J. Imaging Sci., 11 (2018), pp. 2165–2204.

[46] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, *Matrix product state representations*, Quantum Inf. Comput., 7 (2007), pp. 401–430.

[47] I. Pižorn and F. Verstraete, *Variational numerical renormalization group: Bridging the gap between nrg and density matrix renormalization group*, Phys. Rev. Lett., 108 (2012), p. 067202.

[48] Z. Qin, C. Jameson, Z. Gong, M. B. Wakin, and Z. Zhu, *Quantum state tomography for matrix product density operators*, IEEE Trans. Inf. Theory, 70 (2024), pp. 5030–5056.

[49] Z. Qin, M. B. Wakin, and Z. Zhu, *Statistical and algorithmic foundations of probing quantum systems with compressive measurements: A review*, arXiv preprint arXiv:2605.27191, (2026).

[50] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, *Symmetric informationally complete quantum measurements*, J. Math. Phys., 45 (2004), pp. 2171–2180.

[51] T. Rohwedder and A. Uschmajew, *On local convergence of alternating schemes for optimization of convex problems in the tensor train format*, SIAM J. Numer. Anal., 51 (2013), pp. 1134–1162.

[52] U. Schollwöck, *The density-matrix renormalization group*, Rev. Mod. Phys., 77 (2005), pp. 259–315.

[53] T. Shi, M. Ruth, and A. Townsend, *Parallel algorithms for computing the tensor-train decomposition*, SIAM J. Sci. Comput., 45 (2023), pp. C101–C130.

[54] S. S. Sofi, C. Vermeylen, and L. De Lathauwer, *Tensor train quantum state tomography using compressed sensing*, in 2025 33rd European Signal Processing Conference (EUSIPCO), 2025, pp. 1332–1336.

[55] S. S. Sofi, C. Vermeylen, and L. De Lathauwer, *Tomography by design: An algebraic approach to low-rank quantum states*, arXiv preprint arXiv:2602.15202, (2026).

[56] X. Tang, H. Chen, Y. Khoo, and L. Ying, *Sketch tomography: Hybridizing classical shadow and matrix product state*, arXiv preprint arXiv:2512.03333, (2025).

[57] S. Tariq, A. Farooq, J. U. Rehman, T. Q. Duong, and H. Shin, *Efficient quantum state estimation with low-rank matrix completion*, EPJ Quantum Technol., 11 (2024), p. 50.

[58] R. T. Thew, K. Nemoto, A. G. White, and W. J. Munro, *Qudit quantum-state tomography*, Phys. Rev. A, 66 (2002), p. 012303.

[59] G. Tóth, W. Wieczorek, D. Gross, R. Krischek, C. Schwemmer, and H. Weinfurter, *Permutationally invariant quantum tomography*, Phys. Rev. Lett., 105 (2010), p. 250403.

[60] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[61] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac, *Matrix product density operators: Simulation of finite-temperature and dissipative systems*, Phys. Rev. Lett., 93 (2004), p. 207204.

[62] Y. Wang, *Asymptotic equivalence of quantum state tomography and noisy matrix completion*, Ann. Stat., 41 (2013), pp. 2462–2504.

[63] S. R. White, *Density-matrix algorithms for quantum renormalization groups*, Phys. Rev. B, 48 (1993), p. 10345.

[64] S. R. White, *Density matrix renormalization group algorithms with a single center site*, Phys. Rev. B, 72 (2005), p. 180403(R).