Tensor Train Quantum State Tomography using Compressed Sensing

Shakir Showkat Sofi
Dept. Electrical Engineering (ESAT)
KU Leuven
Kortrijk, Belgium
shakirshowkat.sofi@kuleuven.be

Charlotte Vermeylen
Dept. Electrical Engineering (ESAT)
KU Leuven
Leuven, Belgium
charlotte.vermeylen@kuleuven.be

Lieven De Lathauwer
Dept. Electrical Engineering (ESAT)
KU Leuven
Kortrijk/Leuven, Belgium
lieven.delathauwer@kuleuven.be

***Abstract*—Quantum state tomography (QST) 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 impractical due to the exponential growth of parameters in the state representation. In this work, we address this challenge by parameterizing the state using a low-rank block tensor train decomposition and demonstrate that our approach is both memory- and computationally efficient. This framework applies to a broad class of quantum states that can be well approximated by low-rank decompositions, including pure states, nearly pure states, and ground states of Hamiltonians.**

***Index Terms*—low-rank approximation, quantum state tomography, tensor completion, tensor train**

I. Introduction

Quantum computing has garnered significant interest in recent years due to emerging tools that enable the analysis of large-scale systems and optimization over high-dimensional spaces. The study of quantum systems is increasingly crucial in both academia and industry, with applications in quantum communication, sensing, and control. A quantum system's state is represented by a density matrix—a Hermitian, positive semidefinite (PSD) matrix with unit trace. Density matrices provide a unified framework for representing both pure and mixed states (probabilistic mixtures of pure states) [1], [2].

Determining the density matrix of a quantum system is essential yet challenging. For instance, how can we verify that a quantum device produces the intended output? Quantum state tomography (QST) is a fundamental tool for benchmarking and verifying quantum devices [1]–[4]. QST estimates an unknown density matrix by measuring an ensemble of identically prepared quantum systems [1]–[3], [5], [6]. It involves two main steps: (i) performing measurements on a large number of identically prepared quantum systems/circuits and (ii) classically post-processing the data to reconstruct the density matrix. However, as the system size grows, the number of parameters in the density matrix increases exponentially—an obstacle known as the "curse of dimensionality."

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, Charlotte Vermeylen, and Lieven De Lathauwer are affiliated with Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

In practice, many quantum states of interest exhibit special properties, such as being nearly pure or corresponding to ground states of local Hamiltonians [5], [6]. In such cases, the density matrix is low-rank, meaning it resides in a lower-dimensional subspace. This allows for more efficient estimation, as fewer measurements may suffice for accurate recovery. By leveraging connections between low-rank matrix completion and QST, convex optimization methods have been proposed to solve semidefinite programs (SDPs) with nuclear norm penalization. These approaches enable density matrix reconstruction with fewer measurements than traditional tomography methods while maintaining robustness to noise [5]–[10]. However, they implicitly use rank information while optimizing over the full matrix, limiting their scalability.

For large-scale SDPs, an efficient heuristic is the Burer-Monteiro factorization [11], which explicitly imposes a low-rank structure. This approach reformulates the problem non-convexly but significantly improves efficiency by optimizing only over low-rank factors. Inspired by this idea, several QST algorithms employing a Cholesky-like low-rank parametrization of the density matrix have been developed [12]–[14].

Another line of research models the density matrix as a high-order tensor and employs tensor networks—contracted networks of low-order tensors—for compression, thereby mitigating the curse of dimensionality. Common tensor network representations include matrix product states/operators (MPS/MPO) or tensor trains (TT) [15]–[18], tree tensor networks, and projected entangled pair states [19]. QST has been shown to be feasible for certain large-scale quantum many-body systems whose states can be approximated by low-rank tensor networks, such as ground states, GHZ states, cluster states, and AKLT states [3], [4], [20]–[22]. The MPS-based QST approach has demonstrated both efficiency and scalability [3], [4], [21]. In [20], QST is generalized for MPOs, extending MPS methods from pure states to noisy mixed states [16]. However, ensuring that the reconstructed density matrix satisfies necessary constraints (Hermitian, PSD, and unit trace) remains a challenge.

In this work, we adopt a block tensor train (Block-TT) [23], [24] parametrization for the density matrix. Unlike other methods, this approach inherently preserves positive semidefiniteness without requiring additional constraints—analogous

to the Cholesky decomposition for matrices. We demonstrate that our method is both memory- and computationally efficient while accurately recovering mixed quantum states.

A. Preliminaries and Notation

We use lower-case, bold lower-case, bold capital, and calligraphic letters to denote scalars, vectors, matrices, and tensors, respectively, i.e., $x, \mathbf{x}, \mathbf{X}$, and $\mathcal{X}$, respectively. 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 Fig. 1. 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 $\|\cdot\|_*$, respectively. A density matrix is denoted by $\boldsymbol{\rho}$, and a Hilbert space by $\mathcal{H}$. The outer product $\mathcal{Z} = \mathbf{X} \otimes \mathbf{Y}$ 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 $\mathbf{Z} = \mathbf{X} \otimes \mathbf{Y}$ of the same matrices satisfies $z_{k+(i-1)J_1, l+(j-1)J_2} = x_{ij} y_{kl}$. A product (contraction) between the last mode of tensor $\mathcal{X} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ and the first mode of tensor $\mathcal{Y} \in \mathbb{C}^{J_1 \times J_2 \times \cdots \times J_M}$, where $I_N = J_1 = K$, yields an $(N + M - 2)$-order tensor $\mathcal{Z} = \mathcal{X} \bullet \mathcal{Y}$, the entries of which are given by:

$$z_{i_1 \ldots i_{N-1} j_2 \ldots j_M} = \sum_{k=1}^{K} x_{i_1 \ldots i_{N-1} k} y_{k j_2 \ldots j_M}.$$

Similarly, a contraction along mode n is represented by $\mathcal{Z} = \mathcal{X} \bullet_n \mathcal{Y}$, where $I_n = J_n$. Contraction of all common modes is denoted by $\mathcal{X} \bar{\bullet} \mathcal{Y}$. For tensors of the same size, this is equivalent to the inner product $\langle \mathcal{X}, \mathcal{Y} \rangle = \mathcal{X} \bar{\bullet} \mathcal{Y}$. We use tensor network diagrams to visualize tensor networks, see Fig. 1.

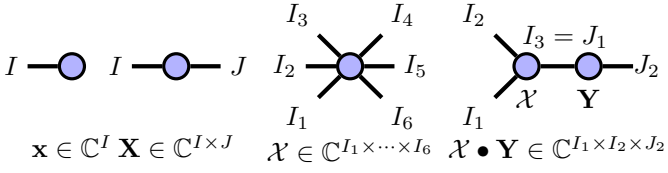

Fig. 1: Basic tensor diagrams.

A *qubit* (quantum bit) is the fundamental unit of quantum information, analogous to a classical bit, but with unique quantum properties. Unlike a classical bit, which can be either 0 or 1, a qubit can exist in a superposition of both states simultaneously. Quantum states can be represented as vectors in a Hilbert space, e.g., for a single qubit, the quantum state is a vector in a two-dimensional Hilbert space. The *Pauli spin matrices* for a single qubit are denoted by $\mathbf{I}_2, \boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y, \boldsymbol{\sigma}_z \in \mathbb{C}^{2 \times 2}$, and are defined by $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, respectively. They are used to measure the state of a quantum system.

A *tensor train matrix (TTM)*, mathematically equivalent to the MPO, is a special type of tensor train decomposition used to represent a large-scale matrix, $\mathbf{X} \in \mathbb{C}^{I \times J}$, as a $2N$th-order tensor, $\mathcal{X} \in \mathbb{C}^{I_1 \times J_1 \times I_2 \times J_2 \times \cdots \times I_N \times J_N}$, where $I = I_1 I_2 \ldots I_N$ and $J = J_1 J_2 \ldots J_N$ [15]–[18]. The TTM decomposition of $\mathcal{X}$ is written as a contraction of a sequence of fourth-order core tensors $\mathcal{G}^{(n)} \in \mathbb{C}^{R_{n-1} \times I_n \times J_n \times R_n} (1 \leq n \leq N)$ with $R_0 = R_N = 1$ such that each entry of $\mathcal{X}$ can be expressed as the sequence of matrix products:

$$x_{i_1 j_1 i_2 j_2 \cdots i_N j_N} = \mathbf{G}^{(1)}_{:i_1 j_1:} \mathbf{G}^{(2)}_{:i_2 j_2:} \cdots \mathbf{G}^{(N)}_{:i_N j_N:}, \tag{1}$$

where matrices $\mathbf{G}^{(n)}_{:ij:} = \mathcal{G}^{(n)}(:, i, j, :) \in \mathbb{C}^{R_{n-1} \times R_n}$ are slices of the core tensor $\mathcal{G}^{(n)}$. The minimal $(R_0, R_1, \ldots, R_N)$ for which Eq. (1) holds is the TT-rank of $\mathcal{X}$, denoted by $\mathbf{r}_{\mathrm{TT}}(\mathcal{X})$. This representation can be much more memory-efficient because it only requires storing the core tensors. The TTM decomposition of a given tensor $\mathcal{X}$ can be computed by a sequence of truncated SVDs (TT-SVD) [17], [18]. We denote the (TT-SVD) operation that projects a given tensor $\mathcal{X}$ to a TTM format, $\left\{\mathcal{G}^{(n)} \in \mathbb{C}^{R_{n-1} \times I_n \times J_n \times R_n}\right\}_{n=1}^{N}$, by $\mathcal{P}_{\mathbf{r}_{\mathrm{TT}}}$. Without loss of generality, we assume that this operation yields a left-orthogonal TT[1]. A visualization of the TTM decomposition is shown in Fig. 2. If all TT-ranks are equal

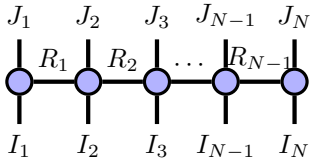

Fig. 2: Tensor diagram of a TTM decomposition.

to 1, then the TTM decomposition is the outer product of N matrices. An N-qubit Pauli matrix can be written as the Kronecker product of N Pauli matrices: $\mathbf{E}_m := \otimes_{j=1}^{N} \boldsymbol{\sigma}_{m_j}$ with $\boldsymbol{\sigma}_{m_j} \in \{\mathbf{I}_2, \boldsymbol{\sigma}_x, \boldsymbol{\sigma}_y, \boldsymbol{\sigma}_z\}$. Such a matrix can be represented in the TTM format by replacing the Kronecker product with the outer product: $\mathcal{E}_m := \otimes_{j=1}^{N} \boldsymbol{\sigma}_{m_j}$.

The Block-n-TT format is a special case of the TTM format, typically used to represent tall matrices, with third-order core tensors except for one fourth-order core at position n [23], [24].

B. Organization

Section II provides a concise overview of compressed sensing QST. Next, in Section III, we introduce our approach to TT QST, together with the algorithm detailed in Section III-B. Finally, we present numerical experiments in Section IV and a brief conclusion in Section V.

II. Compressed sensing quantum state tomography

Consider an N-qubit quantum system with a density matrix $\boldsymbol{\rho} \in \mathbb{C}^{D \times D}$, where $D = 2^N$. According to the quantum theory, when we measure an observable $\mathbf{E}_m \in \mathbb{C}^{D \times D}$, which is a Hermitian matrix, the measurement outcome is one of its (real) eigenvalues. However, we do not know which eigenvalue from the spectrum will pop up. The determinism in quantum mechanics is only of a statistical character, except for certain laws

[1]All the TT cores to the left of $\mathcal{G}^{(N)}$ are left-orthogonal [17], [25].

[2]. The expected measurement outcome can be determined and is given by $y_m = \mathrm{Tr}(\boldsymbol{\rho}\mathbf{E}_m)$. QST estimates the unknown density matrix by performing measurements on an ensemble of identically prepared quantum systems (i.e., prepared in the same quantum state). In estimation terminology, given (possibly noisy) measurement data $\{\mathbf{E}_m, y_m\}_{m=1}^M$, we are required to estimate a density matrix $\hat{\boldsymbol{\rho}}$ that agrees with the observed data, i.e., $y_m \approx \mathrm{Tr}(\hat{\boldsymbol{\rho}}\mathbf{E}_m),$ for $m = 1, 2, \ldots, M$. The number of measurements required to ensure the unique recovery of the (full-rank) density matrix is approximately $\mathcal{O}(D^2)$. However, in the rank-R case, $\mathcal{O}(RD\log^2(D))$ measurements suffice to uniquely recover the density matrix with high probability [5]–[10]. The rank minimization of QST (with nuclear norm relaxation of the rank) can be expressed as:

$$\begin{aligned} &\min_{\hat{\boldsymbol{\rho}}} \|\hat{\boldsymbol{\rho}}\|_* \quad \text{s.t.} \quad \|\mathbf{y} - \mathcal{M}(\hat{\boldsymbol{\rho}})\|_2 \le \epsilon, \text{ and } \hat{\boldsymbol{\rho}} \in \mathcal{S}, \\ &\text{where } \mathcal{S} = \{\mathbf{X} \mid \mathbf{X} \succeq 0,\ \mathrm{Tr}(\mathbf{X}) = 1,\ \mathbf{X} = \mathbf{X}^{\mathrm{H}}\}, \end{aligned} \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^M$ is the vector of measurement outcomes and the measurement map $\mathcal{M}$ is defined such that $(\mathcal{M}(\mathbf{X}))_m = \mathrm{Tr}(\mathbf{E}_m\mathbf{X})$. When N is small, the problem (2) can be easily solved using software such as SDPT3 or CVX(PY) [26]. For large-scale systems, optimizing over a full density matrix becomes expensive. Alternative approaches have been explored that explicitly impose a low-rank decomposition model and optimize over low-rank factors. E.g., it has been shown that one can recast the problem (2) into a problem that implicitly accounts for the Hermitian and PSD constraints by setting $\hat{\boldsymbol{\rho}} = \mathbf{A}\mathbf{A}^{\mathrm{H}}$, where $\mathbf{A} \in \mathbb{C}^{D\times R}$; see, e.g., [11]. In general, the SDP of this parametrization can have local solutions[2]. However, the extra unit trace constraint makes it suitable for many tasks, such as QST [10]–[14]. More recently, projected (factored) gradient descent methods have been proposed with further relaxation of the unit trace by a convex constraint $\|\mathbf{A}\|_{\mathrm{F}}^2 \le 1 \Leftrightarrow \mathrm{Tr}(\hat{\boldsymbol{\rho}}) \le 1$ [10], [13]. Our approach generalizes such parametrizations to tensor networks.

III. Tensor Train Quantum State Tomography

As noted earlier, the MPS (TT) efficiently represents a range of quantum many-body (pure) states and comes with more advanced algorithms, known as density matrix renormalization group algorithms (DMRG), for solving one-dimensional quantum systems [15], [27], [28]. The MPO (TTM) extend the MPS to mixed states [16]. In this work, we employ a representation that expresses a density matrix as a contraction of two Block-TT networks. This parametrization allows for representing a mixed state with asymptotically the same number of parameters as that of a TT[3]. Our experiments show that this representation does not require additional constraints to ensure positive semi-definiteness. In this representation, a $2N$th-order density tensor $\boldsymbol{\rho} \in \mathbb{C}^{2\times 2\times\cdots\times 2}$ is expressed as a contraction of an $(N+1)$th-order Block-TT $\mathcal{A}$ with its Hermitian transpose $\mathcal{A}^{\mathrm{H}}$, i.e., $\boldsymbol{\rho}_{\mathrm{TT}} = \mathcal{A} \bullet_n \mathcal{A}^{\mathrm{H}}$. Intuitively, this is similar to the Burer-Monteiro factorization for matrices [11]. A tensor network diagram of this model is shown in Fig. 3. It is interesting to note that the position of the 4th-order core carrying the index K can be moved to other locations. It has been shown that it serves as an extra tuning parameter during optimization [23], [24]. Furthermore, this can be utilized to adjust ranks dynamically: the 4th-order core can be merged with one of its neighbours, and subsequently, this combined core can be decomposed using an SVD of a different rank, which also allows the index K to be shifted to an adjacent position.

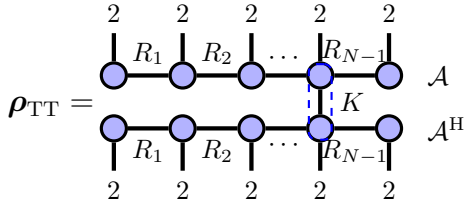

Fig. 3: Decomposition of $\boldsymbol{\rho}_{\mathrm{TT}}$ into Block-TT format, i.e., $\boldsymbol{\rho}_{\mathrm{TT}} = \mathcal{A} \bullet_{(N-1)} \mathcal{A}^{\mathrm{H}}$, where $\mathcal{A} \in \mathbb{C}^{2\times 2\times\cdots\times K\times 2}$.

Although there are different choices for observables (measurements), we use N-qubit Pauli matrices in the TTM format $\{\mathcal{E}_m\}_{m=1}^M$, as they can be decomposed with TT-ranks equal to 1. We utilize the TT algebra to efficiently compute expectation values through a core-wise contraction, as shown in Fig. 4, representing the inner product between $\boldsymbol{\rho}_{\mathrm{TT}}$ and $\mathcal{E}_m$; see, e.g., [17], [28], [29] for details.

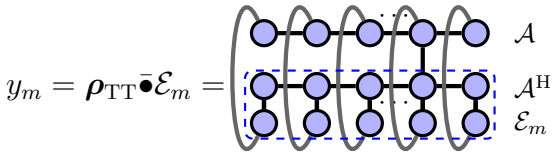

Fig. 4: Expectation values are efficiently calculated by a tensor network contraction. First, $\mathcal{A}^{\mathrm{H}}$ is contracted with $\mathcal{E}_m$ along common modes (shown in the area in the dashed line, i.e., $\mathcal{A}^{\mathrm{H}} \bar{\bullet} \mathcal{E}_m$), and then the resulting tensor is contracted with $\mathcal{A}$ along all the modes.

A. Optimization objective

We recast the problem (2), in line with [10]–[14], as a non-convex problem:

$$\begin{aligned} \min_{\mathcal{A}} \quad & \underbrace{\frac{1}{2}\|\mathbf{y} - \mathcal{M}\left(\hat{\boldsymbol{\rho}}_{\mathrm{TT}}\right)\|_2^2}_{=:f(\mathcal{A})}, \quad \text{with } \hat{\boldsymbol{\rho}}_{\mathrm{TT}} = \mathcal{A} \bullet_n \mathcal{A}^{\mathrm{H}}. \\ \text{s.t.} \quad & \|\mathcal{A}\|_{\mathrm{F}}^2 \le 1, \end{aligned} \tag{3}$$

where the entries of $\mathcal{A}$ are given by $a_{i_1 i_2 \cdots i_n k_n \cdots i_N} = \mathbf{G}^{(1)}_{:i_1:}\mathbf{G}^{(2)}_{:i_2:}\cdots\mathbf{G}^{(n)}_{:i_n,k_n:}\cdots\mathbf{G}^{(N)}_{:i_N:}$. We have reformulated the problem for two reasons: firstly, the number of parameters in the proposed model scales logarithmically with the size of the density matrix[3]. Secondly, this allows us to utilize

[2]Without any additional constraints beyond the SDP 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 $\mathbf{L}\mathbf{L}^{\mathrm{H}} = \mathbf{I}$.

[3]The number of parameters is equal to $2R^2(\log_2 D - 3) + 2R^2K + 4R$, assuming $R_n = R$.

tensor network operations to perform computations efficiently. It is important to note that this approach is efficient only when the quantum states can be well approximated with low TT-rank. The gradient of the objective (3), given by $\nabla f(\mathcal{A}) = -2\sum_{m=1}^{M}(y_m - \hat{\boldsymbol{\rho}}_{\mathrm{TT}} \bar{\bullet} \bar{\mathcal{E}}_m)\,\mathcal{E}_m \bar{\bullet} \mathcal{A}$, can be computed core-wise.

B. Algorithm

This section provides a simple gradient descent (GD) algorithm for optimizing the objective (3). The key steps of the algorithm are as follows. A TT-SVD operation $\mathcal{P}_{\mathbf{r}_{\mathrm{TT}}}(\mathcal{A})$ is performed, which projects $\mathcal{A}$ to a left-orthogonalized Block-TT format. As noted earlier, this process involves truncation and orthonormalization. Truncation (or rounding) is required because the TT-ranks of the gradient and the current iterate get added following the GD step. Alternative approaches that optimize over fixed-rank TT manifolds (see [25], [30]–[32]) could be more suitable; however, here we provide a simple working algorithm. As $\mathcal{A}$ is left-orthogonal, the Frobenius norm of $\mathcal{A}$ is equal to the norm of its last core, $\mathcal{G}^{(N)}$. In the projection step, we normalize $\mathcal{G}^{(N)}$ to ensure that $\|\mathcal{A}\|_{\mathrm{F}}^2 \leq 1$. We noticed that adding a small momentum factor η in the GD step improved the convergence. The pseudocode for the proposed method is outlined in Algorithm 1.

Algorithm 1: Block-TT QST Algorithm.

Data: Gradient of objective $\nabla f(\mathcal{A})$, η, $\mathbf{r}_{\mathrm{TT}}$.
Result: $\mathcal{A}$, such that $\hat{\boldsymbol{\rho}}_{\mathrm{TT}} = \mathcal{A} \bullet_{(N-1)} \mathcal{A}^{\mathrm{H}}$.
Initialize $\mathcal{A}_i$.
$\mathcal{A}_i = \mathcal{P}_{\mathbf{r}_{\mathrm{TT}}}(\mathcal{A}_i)$. % $\mathcal{A}_i$ is left-orthogonal
Projection: $\mathcal{A}_i = \frac{1}{\|\mathcal{G}^{(N)}\|_{\mathrm{F}}}\mathcal{A}_i$. % $\|\mathcal{A}\|_{\mathrm{F}}^2 \leq 1$
for $i = 1 \ldots$ **do**
  Gradient Descent: $\mathcal{A}_{i+1} = \mathcal{P}_{\mathbf{r}_{\mathrm{TT}}}(\mathcal{A}_i - \eta \nabla f(\mathcal{A}_i))$.
  Projection: $\mathcal{A}_{i+1} = \frac{1}{\|\mathcal{G}^{(N)}\|_{\mathrm{F}}}\mathcal{A}_{i+1}$.
end

IV. Experiments

A. Evaluation Metrics

Two common metrics are used to assess performance: (1) Fidelity, which quantifies the closeness between two density matrices $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$: $\left(\mathrm{Tr}\left(\sqrt{\sqrt{\boldsymbol{\rho}_1}\boldsymbol{\rho}_2\sqrt{\boldsymbol{\rho}_1}}\right)\right)^2$ and, (2) Trace distance, which quantifies distinguishability: $\frac{1}{2}\|\boldsymbol{\rho}_1 - \boldsymbol{\rho}_2\|_1$.

B. Results

An 8th-order tensor in Block-TT format $\mathcal{A} \in \mathbb{C}^{2\times 2\times \cdots \times K \times 2}$ is generated with $K = 2$ and $\mathbf{r}_{\mathrm{TT}}(\mathcal{A}) = (1, 2, \ldots, 2, 1)$, by sampling entries from a normal distribution (the Ginibre ensemble). The tensor is normalized and a 14th-order tensor $\boldsymbol{\rho}_{\mathrm{TT}} = \mathcal{A} \bullet_{(N-1)} \mathcal{A}^{\mathrm{H}}$ is computed. Random Pauli measurements are performed with sampling ratio M/D^2 varying from 0.05 to 0.4. Gaussian noise is added to the measurements with SNR $= 60\,\mathrm{dB}$. Algorithm 1 is compared with state-of-the-art algorithms in the reconstruction of $\hat{\boldsymbol{\rho}}_{\mathrm{TT}}$, including a convex optimization method (CVX) [5], [26], maximum likelihood estimation (MLE) [33], and a low-rank projected factor gradient method (LR) [13]. The LR algorithm was initialized randomly. The median Fidelity and Trace distance are shown across 20 trials in Fig. 5.

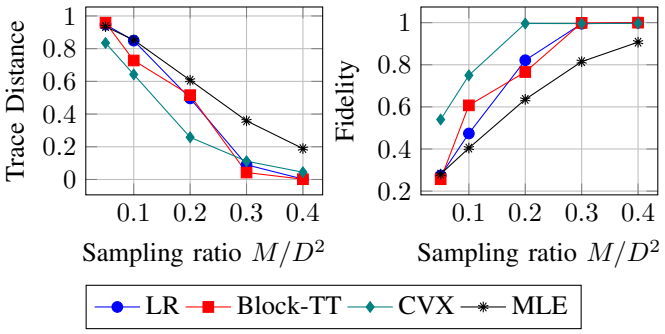

Fig. 5: The proposed method outperforms MLE for higher sampling ratios and has an overall similar performance as LR. The CVX method performs best at low sampling ratios, but this approach solves the original problem (2) and becomes very expensive as the problem size increases.

Next, the number of qubits N is varied from 3 to 12. Random Pauli measurements are performed in which the sampling ratio, M/D^2, was set to 0.05. Fig. 6 shows the median time required to perform M measurements, i.e., $\{\boldsymbol{\rho}_{\mathrm{TT}} \bar{\bullet} \bar{\mathcal{E}}_m\}_{m=1}^{M}$, across 10 trials. This is the most computationally expensive step in the evaluation of the cost function and gradient. In the matrix case, performing M measurement, i.e., $\{\mathrm{Tr}(\boldsymbol{\rho}\mathbf{E}_m)\}_{m=1}^{M}$, becomes exceedingly expensive as N increases, while the proposed format allows performing measurements for large-scale systems.

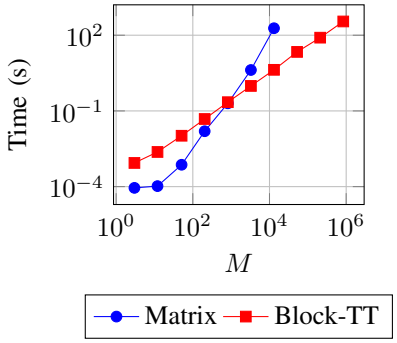

Fig. 6: The time required to perform M measurements in the Block-TT format scales close to linearly, whereas in a matrix format, it scales exponentially.

V. Conclusion and Future Work

A novel QST approach is introduced in which a density matrix is parameterized by a Block-TT network. The proposed method is both memory- and computationally efficient. Numerical experiments showed that the proposed approach yields accurate and valid results without additional constraints. Further theoretical analysis, e.g., of the recovery guarantees, can be useful (some theoretical results of a related approach are known [34]). More efficient algorithms can be designed, for example, by optimizing over TT manifolds [25], [30]–[32].

Furthermore, structured tensor completion approaches could be utilized to perform QST algebraically with deterministic recovery guarantees; see, e.g., [35]–[38].

REFERENCES

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

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

[3] M. Cramer, M. B. Plenio, S. T. Flammia, *et al.*, "Efficient quantum state tomography," *Nat. Commun.*, vol. 1, no. 1, p. 149, 2010.

[4] B. P. Lanyon, C. Maier, M. Holzäpfel, *et al.*, "Efficient tomography of a quantum many-body system," *Nat. Phys.*, vol. 13, no. 12, pp. 1158–1162, 2017.

[5] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150 401, 15 Oct. 2010.

[6] 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.*, vol. 108, p. 170 403, 17 2012.

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

[8] Y. Wang, "Asymptotic equivalence of quantum state tomography and noisy matrix completion," *Ann. Stat.*, vol. 41, no. 5, pp. 2462–2504, 2013.

[9] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, pp. 111–119, 2012.

[10] A. Kalev, R. L. Kosut, and I. H. Deutsch, "Quantum tomography protocols with positivity are compressed sensing protocols," *npj Quantum Inf.*, vol. 1, no. 1, pp. 1–6, 2015.

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

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

[13] 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.*, vol. 4, no. 1, p. 36, 2018.

[14] J. L. Kim, G. Kollias, A. Kalev, K. X. Wei, and A. Kyrillidis, "Fast quantum state reconstruction via accelerated non-convex programming," *Photonics*, vol. 10, no. 2, 2023.

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

[16] F. Verstraete, J. J. García-Ripoll, and J. I. Cirac, "Matrix product density operators: Simulation of finite-temperature and dissipative systems," *Phys. Rev. Lett.*, vol. 93, p. 207 204, 20 2004.

[17] I. Oseledets, "Tensor-train decomposition," *SIAM J. Sci. Comput.*, vol. 33, pp. 2295–2317, 2011.

[18] I. V. Oseledets, "Approximation of $2^d \times 2^d$ matrices using tensor decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 4, pp. 2130–2145, 2010.

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

[20] Z. Qin, C. Jameson, Z. Gong, M. B. Wakin, and Z. Zhu, "Quantum state tomography for matrix product density operators," *IEEE Trans. Inf. Theory*, vol. 70, no. 7, pp. 5030–5056, 2024.

[21] 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*, vol. 109, p. 052 616, 5 2024.

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

[23] S. V. Dolgov, B. N. Khoromskij, I. V. Oseledets, and D. V. Savostyanov, "Computation of extreme eigenvalues in higher dimensions using block tensor train format," *Comput. Phys. Commun.*, vol. 185, no. 4, pp. 1207–1216, 2014.

[24] 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.*, vol. 36, no. 3, pp. 994–1014, 2015.

[25] M. Steinlechner, "Riemannian optimization for high-dimensional tensor completion," *SIAM J. Sci. Comput*, vol. 38, no. 5, S461–S484, 2016.

[26] S. Diamond and S. Boyd, "Cvxpy: A python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, Mar. 2016.

[27] U. Schollwöck, "The density-matrix renormalization group," *Rev. Mod. Phys.*, vol. 77, no. 1, pp. 259–315, 2005.

[28] I. Oseledets, "Dmrg approach to fast linear algebra in the tt-format," *Comput. Meth. Appl. Mat.*, vol. 11, no. 3, pp. 382–393, 2011.

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

[30] S. Holtz, T. Rohwedder, and R. Schneider, "On manifolds of tensors of fixed tt-rank," *Numerische Mathematik*, vol. 120, no. 4, pp. 701–731, 2012.

[31] C. Lubich, I. V. Oseledets, and B. Vandereycken, "Time integration of tensor trains," *SIAM J. Numer. Anal.*, vol. 53, no. 2, pp. 917–941, 2015.

[32] C. Vermeylen and M. Van Barel, "A riemannian rank-adaptive method for higher-order tensor completion in the tensor-train format," *Numer. Linear Algebra Appl.*, vol. 32, no. 1, e2606, 2025.

[33] K. Banaszek, G. D'ariano, M. Paris, and M. Sacchi, "Maximum-likelihood estimation of the density matrix," *Phys. Rev. A*, vol. 61, no. 1, p. 010 304, 1999.

[34] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra Appl.*, vol. 523, pp. 220–262, 2017.

[35] N. Vervliet, O. Debals, L. Sorber, and L. De Lathauwer, "Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in big data analysis," *IEEE Signal Process. Mag.*, vol. 31, pp. 71–79, 2014.

[36] M. Sørensen and L. De Lathauwer, "Fiber sampling approach to canonical polyadic decomposition and application to tensor completion," *SIAM J. Matrix Anal. Appl.*, vol. 40, no. 3, pp. 888–917, 2019.

[37] M. Sørensen, S. Hendrikx, and L. De Lathauwer, "Multilinear singular value decomposition based completion with fibers observed in a single mode," *SIAM J. Matrix Anal. Appl.*, 2025, [Accepted for publication].

[38] S. S. Sofi, S. Hendrikx, and L. De Lathauwer, "Tensor train completion of multi-way data observed along one mode," in *Proceedings of the 32nd EUSIPCO*, 2024, pp. 1067–1071.