All Aboard the Tensor Train

Mike Wakin
Colorado School of Mines

Joint work with:

Zhen Qin, Zhihui Zhu (Ohio State)
Alex Lidiak, Casey Jameson, Alireza Goldar, Zhexuan Gong (Mines)
Gongguo Tang (CU)

Tensor setup and motivation

Order-N tensor $\mathcal{X} \in \mathbb{C}^{d \times d \times \cdots \times d}$ with entries indexed as $\mathcal{X}(i_1, i_2, \ldots, i_N)$.

(Can generalize to dimensions $d_1 \times d_2 \times \cdots \times d_N$.)

Total of d^N entries; scales exponentially in N.

Our interest: scenarios with N large, such as quantum state tomography.

Themes: how to efficiently compute, store, and measure such tensors, avoiding exponential scaling where possible?

Tensor decompositions

Structured models may allow natural or economical parameterizations.

Canonical Polyadic (CP) decomposition: sum of tensor products of rank-one factors.

- Easy to store: $O(N)$ parameters.
- Hard to compute: determining rank and decomposition are NP-hard.

Tucker decomposition: core tensor and set of matrices.

- Easy to compute: computed via higher-order SVD (HOSVD).
- Hard to store: number of parameters is exponential in N.

Tensor trains

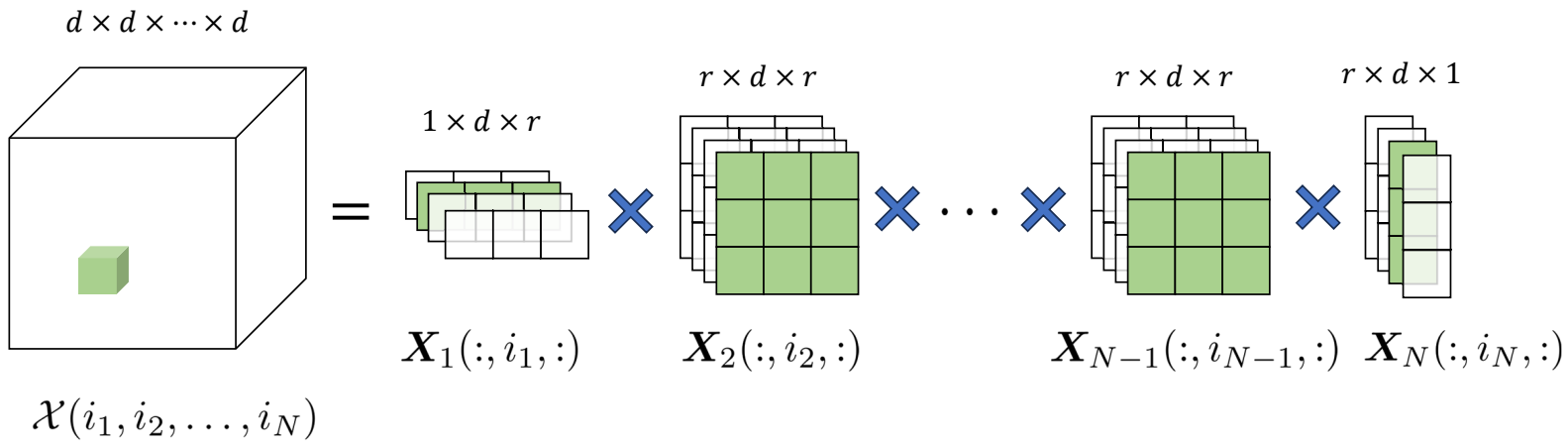
The tensor train (TT) decomposition of $\mathcal{X}$ expresses its entries as

$$\mathcal{X}(i_1, i_2, \ldots, i_N) = \mathbf{X}_1(:, i_1, :)\mathbf{X}_2(:, i_2, :) \cdots \mathbf{X}_N(:, i_N, :)$$

using a collection of third-order tensors

$$\mathbf{X}_1 \in \mathbb{C}^{1 \times d \times r}, \qquad \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_{N-1} \in \mathbb{C}^{r \times d \times r}, \qquad \mathbf{X}_N \in \mathbb{C}^{r \times d \times 1}.$$

Equivalently, each entry of $\mathcal{X}$ is expressed as a *matrix product*



Number of tensor entries: d^N. Number of TT parameters: $O(Ndr^2)$.

Tensor trains - 2

Applications of tensor trains include:

- probabilistic graphical models
- compactly representing large-scale linear operators, such as in deep networks
- image compression
- recommendation systems
- language modeling
- representing states of quantum many-body systems

Not invariant to dimensional permutation.

Closely related: tensor rings, tensor networks.

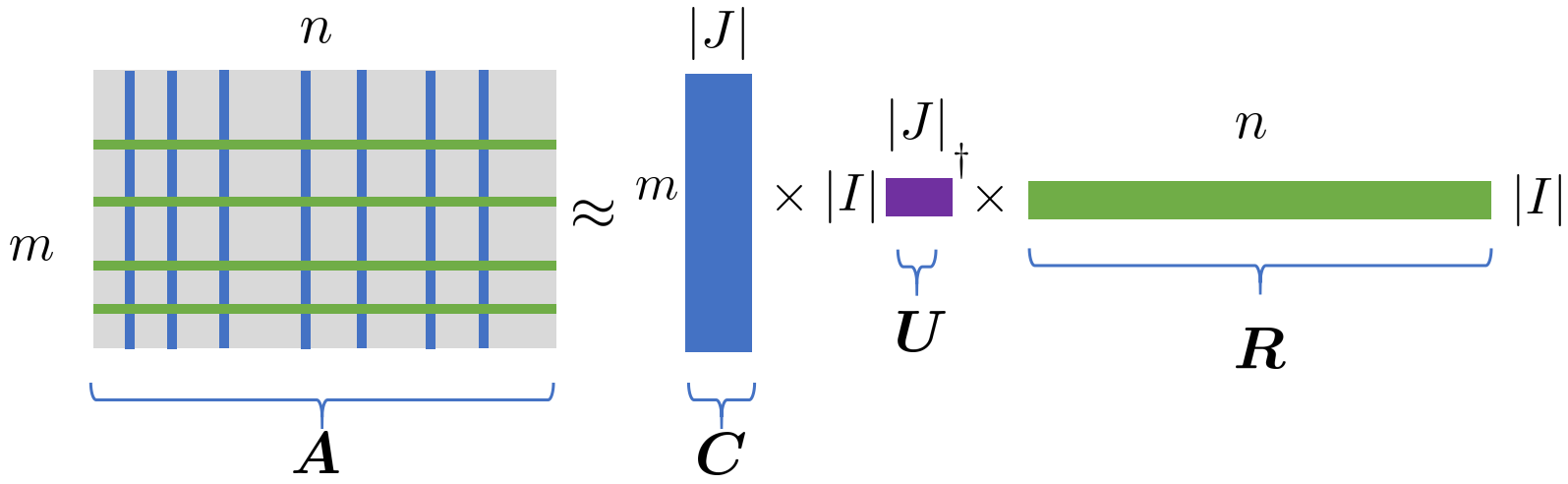
Tensor trains: TT-SVD

A sequential SVD-based algorithm known as tensor train SVD (TT-SVD) (Oseledets, 2011) yields a quasi-optimal TT decomposition.

However, just as a classical SVD requires access to all of the entries of a matrix, the TT-SVD method requires access to all of the entries in $\mathcal{X}$.

Matrix cross approximation

Factor a low-rank matrix using samples from that matrix (Goreinov, 2001):



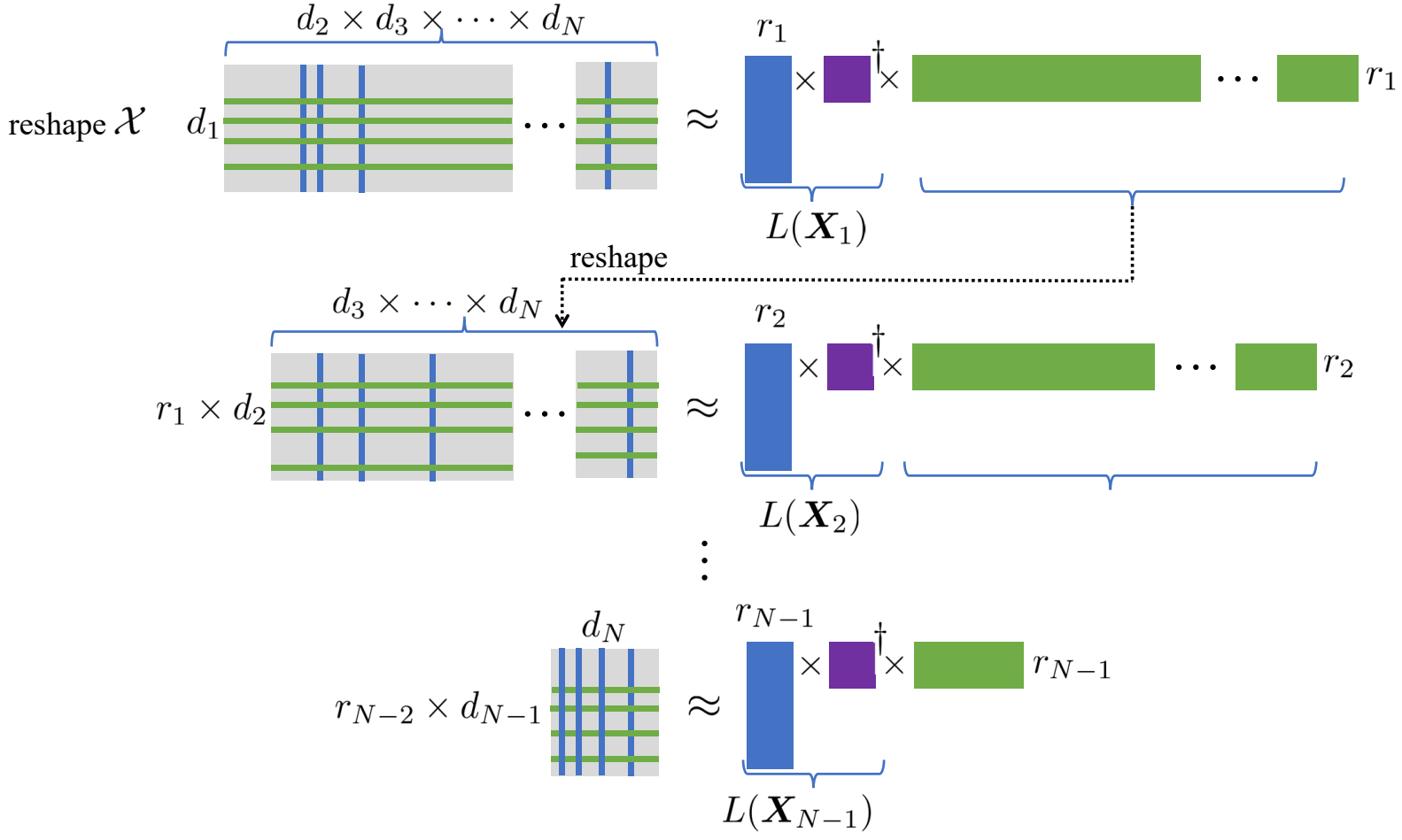
Select row indices $I \subseteq \{1, 2, \ldots, m\}$ and column indices $J \subseteq \{1, 2, \ldots, n\}$.

Define $\mathbf{C} = \mathbf{A}(:, J)$, $\mathbf{U} = \mathbf{A}(I, J)$, $\mathbf{R} = \mathbf{A}(I, :)$.

If $\mathrm{rank}(\mathbf{U}) = \mathrm{rank}(\mathbf{A})$, then $\mathbf{A} = \mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}$.

Tensor train cross approximation

For a tensor $\mathcal{X} \in \mathbb{C}^{d_1 \times d_2 \times \cdots \times d_N}$, the cross approximation process can be generalized to (Oseledets, 2010):



Tensor train cross approximation - 2

General form:

$$\widehat{\mathcal{X}}(i_1, \ldots, i_N) = \prod_{k=1}^{N} \mathbf{X}^{\langle k \rangle}(I^{\leq k-1}, i_k, I^{>k})[\mathbf{X}^{\langle k \rangle}(I^{\leq k}, I^{>k})]^{\dagger}_{\tau_k}$$

with samples of k-th unfolding

$$\mathbf{X}^{\langle k \rangle} \in \mathbb{C}^{(d_1 \cdots d_k) \times (d_{k+1} \cdots d_N)}$$

indexed by interpolation sets $\{I^{\leq k}, I^{>k}\}$

- $I^{\leq k}$ selected from $\{1, 2, \ldots, d_1 \cdots d_k\}$
- $I^{>k}$ selected from $\{1, 2, \ldots, d_{k+1} \cdots d_N\}$
- $I^{\leq 0} = I^{>N} = \emptyset$.

Greedy restricted cross interpolation (GRCI) algorithm works well for choosing interpolation sets.

Tensor train cross approximation - 3

For a tensor $\mathcal{X}$ and its tensor train cross approximation $\widehat{\mathcal{X}}$, an elementwise approximation guarantee has been established (Savostyanov, 2014; Osinsky, 2019):

$$\max_{i_1,\ldots,i_N} |\mathcal{X}(i_1,\ldots,i_N) - \widehat{\mathcal{X}}(i_1,\ldots,i_N)| \leq a^{\lceil \log_2 N \rceil} b,$$

where a and b are constants.

Notably, the guarantee is not exponential in N.

But naively extending the bound to the entire tensor yields

$$\|\mathcal{X} - \widehat{\mathcal{X}}\|_F \leq \sqrt{d_1 \cdots d_N} a^{\lceil \log_2 N \rceil} b,$$

which does scale exponentially.

Tensor train cross approximation - 4

Theorem

Suppose $\mathcal{X}$ can be approximated by a TT format tensor with rank $r = \max_{k=1,\ldots,N-1} r_k$ and approximation error

$$\epsilon := \max_{k=1,\ldots,N-1} \|\mathbf{X}^{\langle k \rangle} - \mathbf{X}^{\langle k \rangle}_{r_k}\|_F.$$

For any interpolation sets $\{I^{\leq k}, I^{>k}\}$ such that $\mathrm{rank}(\mathbf{X}^{\langle k \rangle}_{r_k}(I^{\leq k}, I^{>k})) = r_k$, $k = 1, \ldots, N-1$, the cross approximation $\widehat{\mathcal{X}}$ with appropriate thresholding parameters τ_k for the truncated pseudo-inverse satisfies

$$\|\mathcal{X} - \widehat{\mathcal{X}}\|_F \lesssim (a^2 r + a^2 c r \epsilon + a^2 c^2 \epsilon^2)^{\lceil \log_2 N \rceil - 1} (a^2 \epsilon + a^2 c \epsilon^2 + a^2 c^2 \epsilon^3),$$

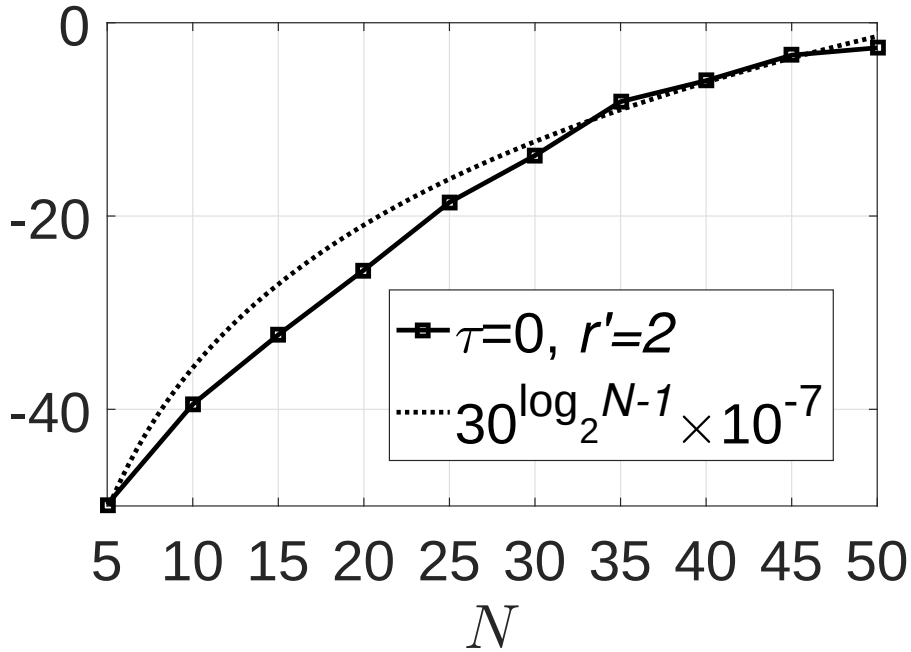
where a and c are constants.

Can extend to noisy samples, absorbing noise level into ϵ.

Tensor train cross approximation - 5

Plot of mean square error in dB:

$$\text{MSE} = 10\log_{10} \|\mathcal{X} - \widehat{\mathcal{X}}\|_F^2 / \|\mathcal{X}\|_F^2$$



Parameters: $d = 2$, $r' = r = 2$, noise $\mu = 10^{-5}$, and low-rank error $\eta = 0$.

Quantum state tomography

The state of an N-qubit quantum system is described by a density matrix $\rho \in \mathbb{C}^{2^N \times 2^N}$ that is PSD and has $\mathrm{trace}(\rho) = 1$.

This matrix has 4^N elements and is naturally associated with a high-order tensor.

For example, the *Pauli matrices* are defined as:

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \sigma_1 = \sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \ \sigma_2 = \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \ \sigma_3 = \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

We can express a density matrix in an orthogonal basis formed by Kronecker products of Pauli matrices:

$$\rho = \sum_{i_1, i_2, \ldots, i_N = 0}^{3} \mathcal{X}(i_1, i_2, \ldots, i_N) \left(\sigma_{i_1} \otimes \sigma_{i_2} \otimes \cdots \otimes \sigma_{i_N}\right).$$

The order-N tensor $\mathcal{X}$ contains the Pauli coefficients of ρ.

Quantum state tomography - 2

Each entry of the tensor $\mathcal{X}(i_1, i_2, \ldots, i_N)$ can be sampled by "measuring" the quantum system in an appropriate measurement basis. (More on that soon.)

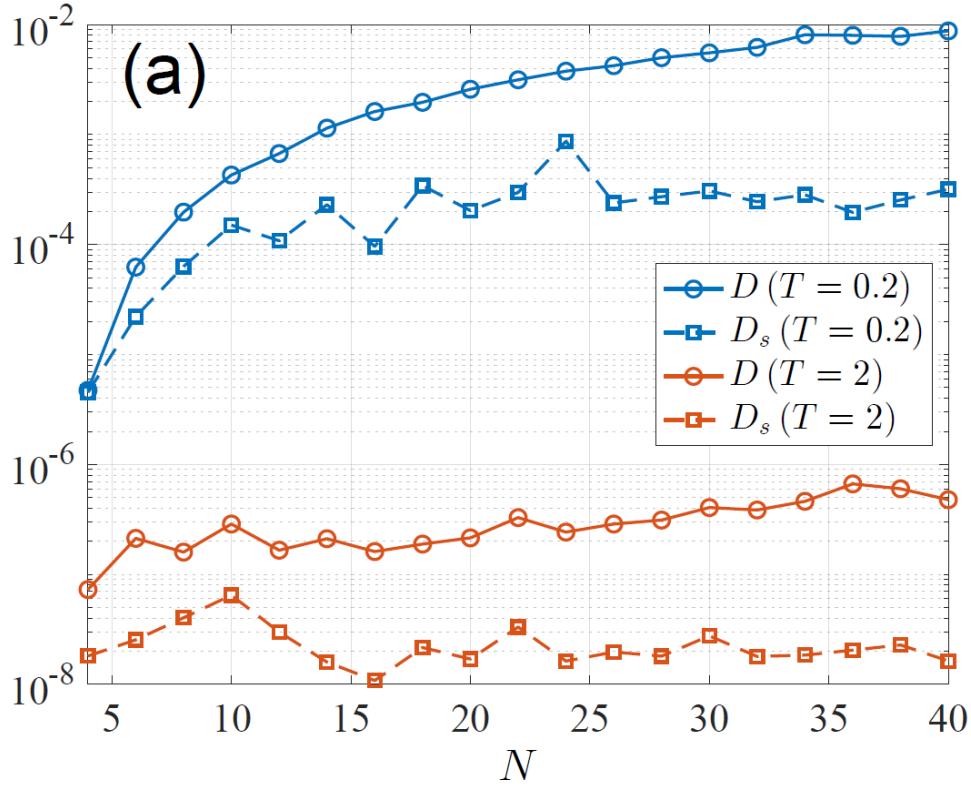
Reconstructing the quantum state from such measurements is known as *quantum state tomography*. However, measuring all 4^N entries is prohibitively expensive.

Fortunately, tensor train models also arise in quantum information contexts, where they are known as *matrix product state (MPS)* and *matrix product operator (MPO)* models. The TT rank r is known as the *bond dimension*.

MPS/MPO models have been shown to describe most states generated by a one-dimensional noisy quantum computer. This presents an opportunity for using TT cross approximation to enable quantum state tomography with a polynomial (in N) number of measurement bases.

Quantum state tomography - 3

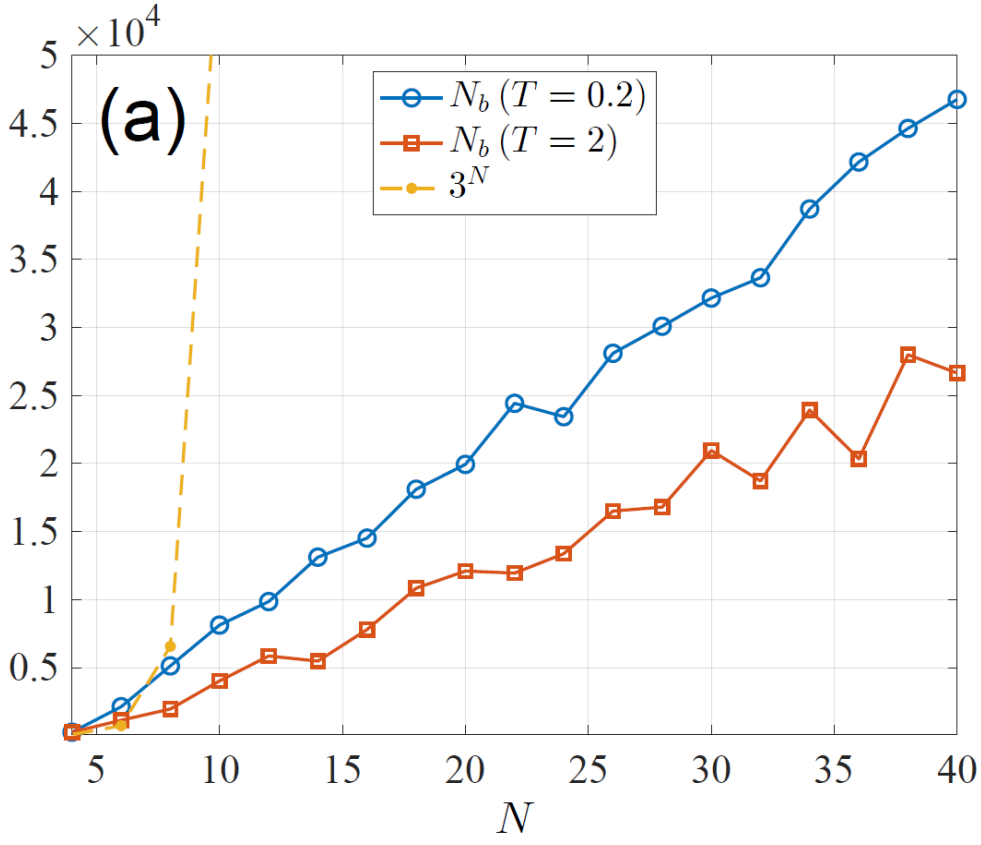
Distance $D = \|\rho - \widehat{\rho}\|_F^2 / \|\rho\|_F^2$; thermal states of 1D quantum Ising model:



Ignores statistical error in measurements; maximum bond dimension used in DMRG-cross is 10, which is suitable for high temperature T but insufficient for low temperature T.

Quantum state tomography - 4

Number of measurement bases required:



Tensor recovery in the TT format

Several works have studied the recovery of tensor train models from incomplete measurements.

Alternating minimization and gradient descent methods for estimating the TT factors lack theoretical guarantees.

Other methods such as iterative thresholding and Riemannian gradient descent require estimating the entire tensor $\mathcal{X}$ at each iteration, and may require additional information or assumptions.

Tensor recovery in the TT format - 2

We study the recovery of a tensor $\mathcal{X}^\star$ from linear measurements

$$\mathbf{y} = \mathcal{A}(\mathcal{X}^\star) \in \mathbb{C}^m.$$

We consider the nonconvex constrained optimization problem

$$\min_{\mathbf{X}_1,\ldots,\mathbf{X}_N} \quad \frac{1}{2m}\|\mathcal{A}([\mathbf{X}_1,\ldots,\mathbf{X}_N]) - \mathbf{y}\|_2^2$$

$$\text{such that } \sum_{i_n=1}^{d_n} \mathbf{X}_n^\top(:,i_n,:)\mathbf{X}_n(:,i_n,:) = \mathbf{I}_{r_n}, n = 1,2,\ldots,N-1,$$

which optimizes the TT model in *left canonical form*.

We implement a (hybrid) Riemannian gradient descent (RGD) algorithm on the TT factors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, projecting back onto the Stiefel manifold at each iteration.

Tensor recovery in the TT format - 3

Theorem

Consider a tensor $\mathcal{X}^\star$ with TT ranks $\mathbf{r} = (r_1, \ldots, r_{N-1})$ and suppose $\mathcal{A}(\cdot) = \mathcal{I}(\cdot)$. Suppose that RGD is initialized with $\{\mathbf{X}_n^{(0)}\}$ satisfying

$$\operatorname{dist}^2(\{\mathbf{X}_n^{(0)}\}, \{\mathbf{X}_n^\star\}) \leq \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{72(N^2-1)(N+1+\sum_{n=2}^{N-1} r_n)}, \tag{1}$$

and the step size $\mu \leq \frac{1}{9N-5}$. Then, the iterates $\{\mathbf{X}_n^{(t)}\}_{t\geq 0}$ generated by RGD will converge linearly to $\{\mathbf{X}_n^\star\}$ (up to rotation):

$$\begin{aligned}&\operatorname{dist}^2(\{\mathbf{X}_n^{(t+1)}\}, \{\mathbf{X}_n^\star\}) \\ &\quad \leq \left(1 - \frac{\underline{\sigma}^2(\mathcal{X}^\star)}{64(N+1+\sum_{n=2}^{N-1} r_n)\|\mathcal{X}^\star\|_F^2}\mu\right) \operatorname{dist}^2(\{\mathbf{X}_n^{(t)}\}, \{\mathbf{X}_n^\star\}).\end{aligned}$$

Note the polynomial dependence on N in the initialization requirement and convergence rate.

Tensor recovery in the TT format - 4

We establish a similar bound (and one with noise) for generic linear maps $\mathcal{A}$ which depends on the RIP constant of $\mathcal{A}$. We also establish a suitable "spectral initialization" using the TT-SVD algorithm:

$$\mathcal{X}^{(0)} = \mathsf{SVD}_{\mathbf{r}}^{\mathsf{TT}}\left(\frac{1}{m}\sum_{k=1}^{m} y_k A_k\right),$$

which is guaranteed to be close to the ground-truth $\mathcal{X}^\star$ when the operator $\mathcal{A}$ satisfies the RIP.

Tensor recovery in the TT format - 5

Convergence rates for tensor sensing:

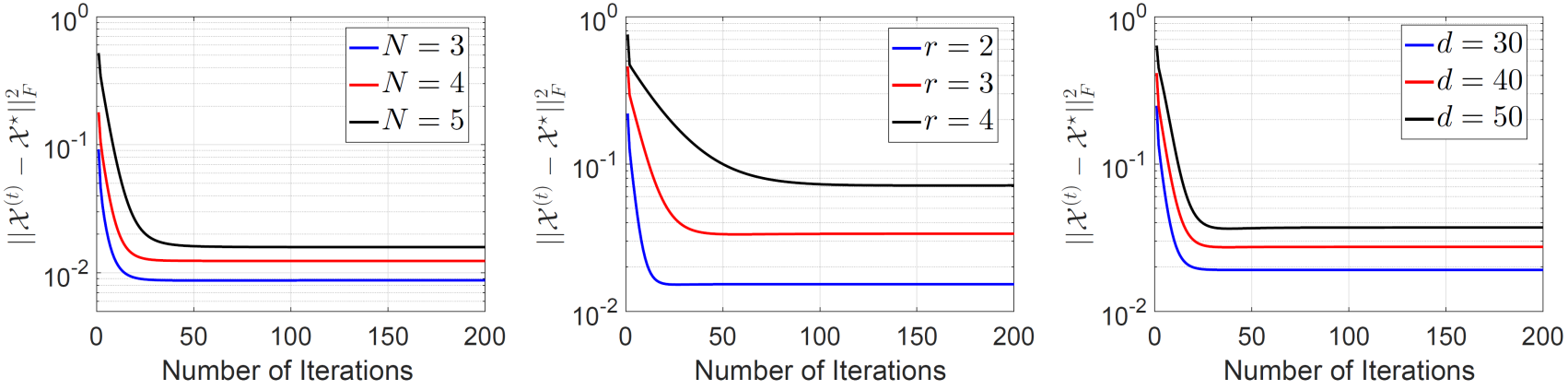


Figure 2: Convergence of RGD for TT format tensor sensing (a) for different N with $d = 10$, $r = 2$, $m = 1000$, and $\gamma^2 = 0.1$, (b) for different r with $d = 50$, $N = 3$, $m = 3000$, and $\gamma^2 = 0.1$, (c) for different d with $N = 3$, $r = 2$, $m = 1500$, and $\gamma^2 = 0.1$.

Gaussian measurements.

Tensor recovery in the TT format - 6

Convergence rates for tensor completion:

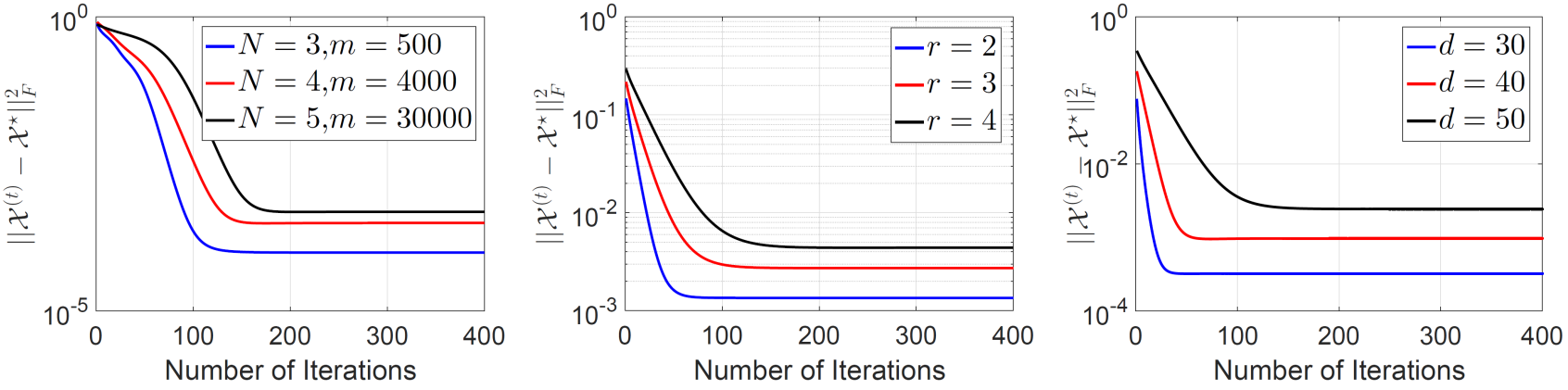


Figure 4: Convergence of RGD for TT format tensor completion (a) for different N and m with $d = 10$, $r = 2$, and $\gamma^2 = 10^{-6}$, (b) for different r with $d = 50$, $N = 3$, $m = 35000$, and $\gamma^2 = 10^{-6}$, (c) for different d with $N = 3$, $r = 2$, $m = 20000$, and $\gamma^2 = 10^{-6}$.

Stable embeddings of tensor trains

We say that a linear measurement operator $\mathcal{A} : \mathbb{C}^{d \times d \times \cdots \times d} \to \mathbb{C}^m$ satisfies the $\delta_{\mathbf{r}}$-restricted isometry property ($\delta_{\mathbf{r}}$-RIP) if

$$(1 - \delta)\|\mathcal{X}\|_F^2 \leq \frac{1}{m}\|\mathcal{A}(\mathcal{X})\|_F^2 \leq (1 + \delta)\|\mathcal{X}\|_F^2$$

for every tensor train $\mathcal{X}$ with ranks up to $\mathbf{r} = (r_1, \ldots, r_{N-1})$.

Extending arguments from Rauhut et al. (2017), we have proved that the $\delta_{\mathbf{r}}$-RIP holds with with high probability for i.i.d. complex Gaussian measurements when

$$m \geq C \frac{1}{\delta_{\mathbf{r}}^2} \cdot N d r^2 \log(Nr),$$

where $r = \max\{r_1, \ldots, r_{N-1}\}$.

Stable embeddings of tensor trains - 2

Our final measurement systems are motivated by quantum state tomography.

Reshaping a $4 \times 4 \times \cdots \times 4$ tensor into a $2^N \times 2^N$ density matrix ρ, our measurements take the form

$$\langle \mathbf{A}_k, \rho \rangle, \qquad k = 1, 2, \ldots, m,$$

where each $\mathbf{A}_k \in \mathbb{C}^{2^N \times 2^N}$.

We consider rank-one measurements of the form $\mathbf{A}_k = \mathbf{a}_k \mathbf{a}_k^H$, where each vector $\mathbf{a}_k \in \mathbb{R}^{2^N}$ is i.i.d. Gaussian. With

$$m \geq CNdr^2 \log N$$

where $d = 4$ for qubits, we obtain the left-half of the RIP bound for every tensor train with ranks up to r.

Stable embeddings of tensor trains - 3

Moving toward more practical measurements, consider a randomly generated Haar-distributed unitary matrix $[\phi_1 \cdots \phi_{2^N}]$. With this we can take up to 2^N rank-one measurements with matrices of the form $\phi_k \phi_k^H$. Aggregating measurements from Q such unitary bases, with

$$Q \geq CNdr^2 \log N$$

where $d = 4$ for qubits, we obtain the left-half of the RIP bound for every tensor train with ranks up to r.

Quantum measurements of tensor trains

When $[\phi_1 \cdots \phi_{2^N}]$ is unitary, the matrices

$$\phi_1\phi_1^H,\ \phi_2\phi_2^H,\ \cdots,\ \phi_{2^N}\phi_{2^N}^H$$

form a Positive Operator-Valued Measure (POVM).

Quantum measurements involving POVMs are not strictly linear; rather than returning

$$p_k = \langle \phi_k\phi_k^H, \rho \rangle, \qquad k = 1, 2, \ldots, 2^N,$$

we collect observations of a random variable whose probability mass distribution (pmf) is given by $p_1, \ldots, p_{2^N}$.

Conducting M total experiments (using M total state copies), we can use the empirical probabilities $\widehat{p}_k$ as approximations of the true p_k.

Quantum measurements of tensor trains - 2

Aggregating measurements from Q such Haar-distributed rank-one POVMs, with M state copies per POVM, any global solution of the constrained least-squares estimator

$$\hat{\rho} = \arg \min_{\rho:\ \mathsf{TT}_r} \|\mathcal{A}(\rho) - \widehat{p}\|_2^2$$

satisfies

$$\|\widehat{\rho} - \rho^*\|_F \leq \epsilon$$

with high probability as long as

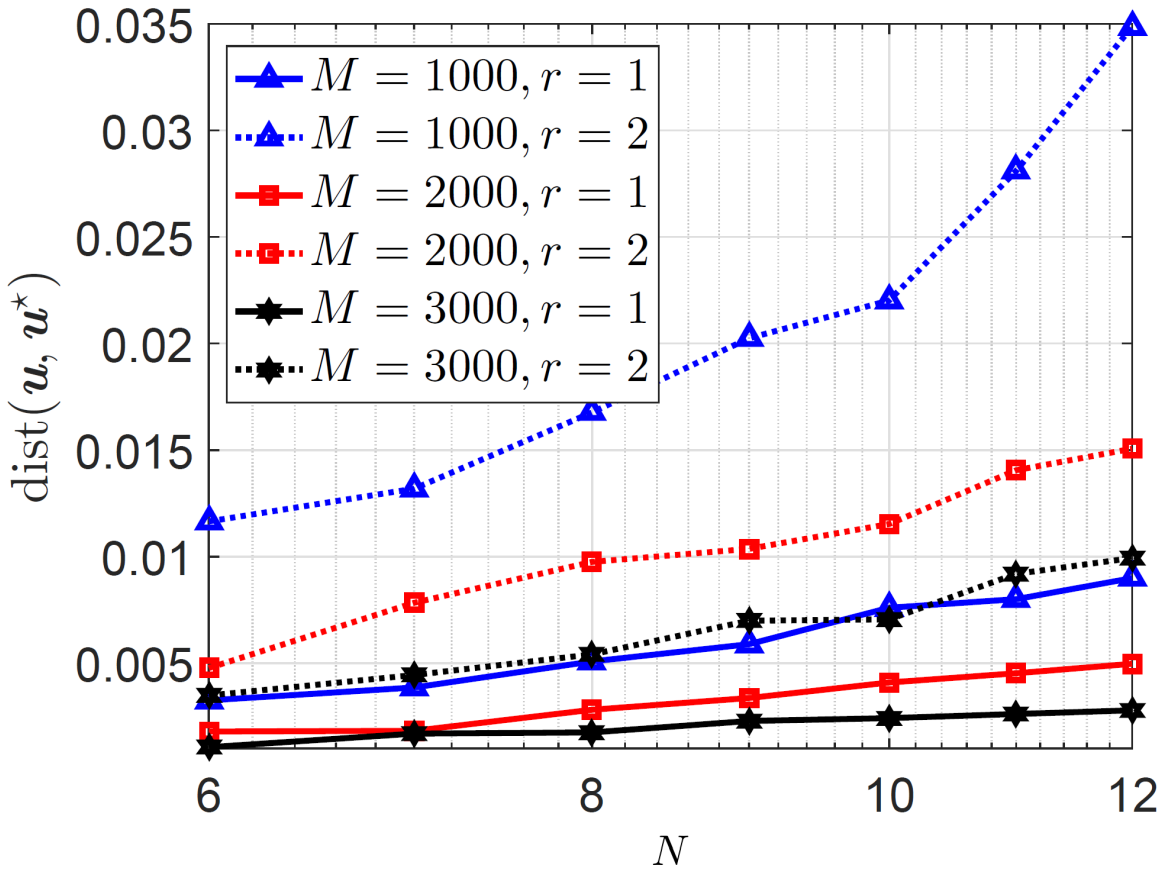
$$QM \geq C \cdot \frac{N^3 d r^2 \,(\text{log factors})}{\epsilon^2}$$

where $d = 4$ for qubits.

This supports the use of low-M measurement schemes.

Quantum measurements of tensor trains - 3

Rank-one MPS; $Q = 1$; iterative hard thresholding.



Error increases with r, decreases with M, increases only polynomially with N.

Conclusions

Tensor trains are a specialized model for representing high-order tensors using a polynomial number of parameters.

When possible, avoiding exponential complexity (in representation, computation, and sampling) facilitates "large N" applications such as quantum state tomography.

References:

- Z. Qin, A. Lidiak, Z. Gong, G. Tang, M. B. Wakin, and Z. Zhu, "Error Analysis of Tensor-Train Cross Approximation," NeurIPS, 2022.
- A. Lidiak, C. Jameson, Z. Qin, G. Tang, M. B. Wakin, Z. Zhu, and Z. Gong, "Quantum state tomography with tensor train cross approximation," arXiv:2207.06397.
- Z. Qin, M. B. Wakin, and Z. Zhu, "Guaranteed Nonconvex Factorization Approach for Tensor Train Recovery," arXiv:2401.02592.
- Z. Qin, C. Jameson, Z. Gong, M. B. Wakin, and Z. Zhu, "Stable Tomography for Structured Quantum States," arXiv:2306.09432. To appear in IEEE Trans. Inform. Theory.