

Tensor Completion with Nearly Linear Samples Given Weak Side Information

CHRISTINA LEE YU, Cornell University, USA

XUMEI XI, Cornell University, USA

Tensor completion exhibits an interesting computational-statistical gap in terms of the number of samples needed to perform tensor estimation. While there are only $\Theta(tn)$ degrees of freedom in a t -order tensor with n^t entries, the best known polynomial time algorithm requires $O(n^{t/2})$ samples in order to guarantee consistent estimation. In this paper, we show that weak side information is sufficient to reduce the sample complexity to $O(n)$. The side information consists of a weight vector for each of the modes which is not orthogonal to any of the latent factors along that mode; this is significantly weaker than assuming noisy knowledge of the subspaces. We provide an algorithm that utilizes this side information to produce a consistent estimator with $O(n^{1+\kappa})$ samples for any small constant $\kappa > 0$. We also provide experiments on both synthetic and real-world datasets that validate our theoretical insights.

CCS Concepts: • **Theory of computation** → **Design and analysis of algorithms**; **Sample complexity and generalization bounds**; • **Mathematics of computing** → *Dimensionality reduction*.

Additional Key Words and Phrases: tensor completion, side information, low rank, matrix estimation

ACM Reference Format:

Christina Lee Yu and Xumei Xi. 2022. Tensor Completion with Nearly Linear Samples Given Weak Side Information. *Proc. ACM Meas. Anal. Comput. Syst.* 6, 2, Article 39 (June 2022), 35 pages. <https://doi.org/10.1145/3530905>

1 INTRODUCTION

A tensor is a mathematical object that can be used to represent multiway data. A dataset in which each datapoint is indexed by t indices can be represented by a t order tensor. A 2-order tensor is simply a matrix, with each datapoint being referenced by two indices referring to the row and column. Multiway data arises in many applications. For example, image data can be represented by a 3-order tensor, with two indices referring to the pixel location, and the third index referring to the color modes of RGB. Video data could then be represented by a 4-order tensor with the 4th mode representing time. E-commerce data is also multiway, with each datapoint of interaction on the platform being associated to a user id, product id, and timestamp. As network data is naturally represented in a matrix form, data collected from monitoring a network changing over time can be represented in a 3-order tensor. Neuroimaging data involves 3D-scans that can be represented in a 3-order tensor. Microbiome studies or protein interaction data involves network data of co-occurrence or interaction counts; this network can be collected across many different patients with different demographics, and it may be useful to represent the full data as a 3-order data to look for patterns amongst subpopulations rather than to just aggregate the data into a single matrix.

Authors' addresses: Christina Lee Yu, Cornell University, Ithaca, New York, USA, cleeyu@cornell.edu; Xumei Xi, Cornell University, Ithaca, New York, USA, xx269@cornell.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

2476-1249/2022/6-ART39 \$15.00

<https://doi.org/10.1145/3530905>

Often the tensor dataset can be very sparse due to the observation process. For example, e-commerce data is very sparse as any given user only interacts with a small subset of the products at sparse timepoints. In experimental studies, each datapoint may be costly to collect, and thus the sparsity could be limited by available budget. The observations themselves can also be noisy or corrupted due to the experimental process or mistakes in data entry. As a result, the task of tensor estimation, or learning the underlying structure given noisy and incomplete tensor data, is a significant building block in the data analysis pipeline.

In the setting of sparse tensor completion, a critical question is how many datapoints does one need to observe (sampled uniformly at random) in order to estimate the underlying tensor structure? Consider a t -order data tensor with n^t entries, i.e. each mode has dimension n . When we only observe a small subset of entries, it is impossible to guarantee recovery without imposing structure on the underlying tensor. The typical assumptions to impose are low rank and incoherence style conditions. Essentially the low rank conditions reduce the number of unknown model parameters to linear in n even though the number of possible tensor entries is n^t for a t -order tensor. The simple statistical lower bound on the sample complexity, or minimum number of observations for recovery, is thus $\Omega(n)$, as there are linear in n degrees of freedom in the low-rank model. Tensor nuclear norm minimization requires $O(n^{3/2})$ observations for a t -order tensor, however the algorithm is not polynomial time computable as tensor nuclear norm is NP-hard to compute [16, 38]. The best existing polynomial time algorithms require $O(n^{t/2})$ observations for a t -order tensor. There is a large gap between what is polynomial time achievable and the statistical lower bound.

For a 3-order tensor, [1] conjectured that $\Omega(n^{3/2})$ samples are needed for polynomial time computation based on a reduction of tensor completion for a class of a rank 1 tensors to the random 3-XOR distinguishability problem. Conditioned on the hardness of random 3-XOR distinguishability, their result proves that any approach for tensor completion that relies on the sum of squares hierarchy or Rademacher complexity will require $\Omega(n^{3/2})$ samples. The fact that the class of hard instances in [1] are simply rank 1 tensors suggests that rank may not be a sufficient measure of complexity for tensor estimation. As a result of the hardness conjecture, recent literature has accepted the threshold of $\Theta(n^{t/2})$ as a likely lower bound for computationally efficient algorithms, and instead has shifted attention to reducing dependence with respect to constant properties of the model, such as rank or incoherence constants.

In this paper we consider what conditions are sufficient to achieve nearly linear sample complexity by the use of auxiliary information. In the most general setting of tensor estimation, the indices of the data entries themselves are not expected to carry valuable information other than linking it to other entries associated to the same index. In particular, the distribution of the data is expected to be equivalent up to permutations of the indices. However, in reality we often have additional knowledge or side information about the indices in each mode such that the full generality of an exchangeable model is not the most appropriate model. For example, with video or imaging data, we expect the data to exhibit smoothness with respect to the pixel location and frame number, which are encoded in the indices of the associated modes. For e-commerce data, there is auxiliary data about the users and products that could relate to the latent factors. While there has been several empirical works studying the potential promise for utilizing side information in tensor completion, none of the works provide theoretical results with statistical guarantees.

1.1 Related Literature

Tensor completion has been studied in the literature as a natural extension of matrix completion to higher dimensions. The approach and techniques have naturally focused around extensions of matrix completion techniques to tensor completion. The earliest approaches unfold the tensor to a matrix,

and apply variations of matrix completion algorithms to the constructed matrix [17, 25, 33, 34]. Given a t -order tensor where each mode has dimension n , there are $2^t - 1$ possible unfoldings of the tensor to a matrix, each corresponding to a partition of the t modes. For example, if τ_A and τ_B are disjoint non-empty subsets of $[t]$ such that $\tau_A \cup \tau_B = [t]$, the corresponding unfolding of the tensor would result in a $n^{|\tau_A|} \times n^{|\tau_B|}$ matrix, where each row would correspond to a member of the cartesian product of coordinates in the modes referenced by τ_A , and each column would correspond to a member of the cartesian product of coordinates in the modes referenced by τ_B . There would still be overall n^t entries in the matrix, with a 1-1 mapping to the n^t entries in the original tensor. The above results utilize the fact that low rank and incoherence conditions for the original tensor result in low rank and incoherence conditions for the unfolding of the tensor to the constructed matrix. As matrix completion algorithms are limited in their sample complexity by the maximum of the dimensions of the rows and columns, the unfoldings that minimize sample complexity are those that unfold the tensor to a $n^{\lfloor t/2 \rfloor} \times n^{\lceil t/2 \rceil}$ matrix, resulting in a sample complexity of $O(n^{\lceil t/2 \rceil})$.

Unfolding the tensor to a matrix is limiting as the algorithm loses knowledge of the relationships amongst the rows and columns of the matrix that were exhibited in the original tensor, i.e. there are rows and columns in the unfolded matrix that share coordinates along some modes of the original tensor. There have subsequently been a series of works that have attempted to use tensor structure to reduce the sample complexity. The majority of results have focused on the setting of a 3rd order tensor, but many of the results can also be extended to general t -order tensors. Tensor nuclear norm minimization requires only sample complexity of $O(n^{3/2})$ for a general t -order tensor, but tensor nuclear norm is NP-hard to compute and thus does not lead to a polynomial time algorithm [16, 37]. The best polynomial time algorithms require a sample complexity of $O(n^{t/2})$ for an order t tensor. These results have been attained using extensions from a variety of techniques similar to the matrix completion setting, including spectral style methods [27, 35], convex relaxation via sum of squares [1, 30], minimizing the nonconvex objective directly via gradient descent [8, 35, 36] or alternating least squares [3, 20], or iterative collaborative filtering [31]. The naive statistical lower bound is $\Omega(n)$ as the number of unknown parameters in a low rank model grows linearly in n . This still leaves a large gap between the sample complexity of the best existing polynomial time algorithms and the statistical lower bound. [1] provides evidence for a computational-statistical gap by relating tensor completion via the Rademacher complexity and sum of squares framework to refutation of random 3-SAT.

While the above works all consider a uniform sampling model, [40] considers an active sampling scheme that achieves optimal sample efficiency of $O(n)$. Their approach requires a specific sampling scheme that aligns all the samples to guarantee that along each subspace there are entire columns of data sampled. While this result yields optimal bounds and is useful for settings where the data can be actively sampled, many applications do not allow such active control over the sampling process. [9] considers the tensor recovery problem, which allows for general measurement operators instead of only single entry observations. They prove that spectral initialization with Riemannian gradient descent can recover the underlying tensor with only $O(nr^2)$ Gaussian measurements, achieving the optimal linear dependence on n . Their result relies on the tensor restricted isometry property that arises from the Gaussian measurements. As such, it does not extend to entrywise observations.

The inductive matrix completion problem considers the setting when exact or partial subspace information is provided alongside the matrix completion task [7, 13, 14, 18, 19]. When exact subspace information is given, the degrees of freedom in the model is significantly reduced as one only needs to estimate the smaller core matrix governing the interaction between the row and column subspaces. As a result, the sample complexity reduces from a linear to logarithmic dependence on the matrix dimension n [19]. [14] further considers the setting where noisy or partial information

about the subspace is given, modeling the desired matrix as a sum of the interaction between the side information and a low rank residual matrix. However, under partial or noisy side information, the sample complexity is still linear in n as the degrees of freedom in the model is still linear in n . There have been empirical works showing benefits of utilizing side information for tensor completion. The most common model assumes that the side information is given in the form of a subspace for each of the modes that contains the associated column subspace of each matricization associated to an unfolding of the tensor [2, 6, 10, 29, 43]; this is a natural extension of the inductive matrix completion setting. Given the subspaces, since the degrees of freedom in the model no longer grows with n , one would expect that the sample complexity would reduce to logarithmic in n , although none of these papers provide formal statistical guarantees. [28] assumes that the side information is given in the form of similarity matrices amongst indices in each mode of the tensor, which they incorporate into the algorithm via a Laplacian regularizer. [23] considers a Bayesian setup in which the side information is in the form of kernel matrices that are used to construct Gaussian process priors for the latent factors. None of these above mentioned results in tensor completion with side information provide formal statistical guarantees, although the empirical results seem promising.

We utilize a similar insight as [22], which shows that for orthogonal symmetric tensors, when all entries are observed, tensor decomposition can be computed by constructing $n \times n$ matrices with the same latent factors as the tensor. Their setting assumes that all entries are observed, and thus does not provide an algorithm or statistical guarantees for noisy and sparsely observed datasets. We extend the idea to sparsely observed tensors and beyond orthogonal symmetric tensors. Our specific algorithm uses the nearest neighbor collaborative filtering approach for matrix completion, introduced in [4, 5, 24, 31, 32].

1.2 Contributions

Consider a t -order tensor with dimension n along each mode, and assume we are given a sparse set of noisy observations where each entry is sampled independently with probability p and observed with mean zero bounded noise. When the tensor has low orthogonal CP-rank, we assume a weak form of side information consisting of a single vector for each mode which simply lies in the column space of the associated matricization of the tensor along that mode. Furthermore suppose the side information vectors are not exactly aligned with any of the latent factors. Under these assumptions, we provide a simple polynomial time algorithm which provably outputs a consistent estimator as long as the number of observed entries is at least $\Omega(n^{1+\kappa})$ for any arbitrarily small constant $\kappa > 0$, nearly achieving the linear lower bound resulting from the degrees of freedom in the model. We extend our results beyond low orthogonal CP-rank tensors as well, providing a characterization for sufficient side information to admit nearly linear sample complexity.

To our knowledge, this is the first theoretical result for tensor completion with side information, provably showing that given weak side information the sample complexity of tensor estimation can reduce from the conjectured $n^{t/2}$ to nearly linear in n . The side information we assume is significantly weaker than assuming knowledge of the full subspaces, and thus is more plausible for real applications. Our proposed algorithm is simple, essentially using matrix estimation techniques on constructed matrices of size $n \times n$ to learn similarities between coordinates. These similarities are used to estimate the underlying tensor via a nearest neighbor estimator. An additional benefit of our analysis is that we are able to prove that with high probability, the maximum entrywise error of our estimate decays as $\tilde{O}(\max(n^{\kappa/4}, n^{-(\kappa+1)/(t+2)})$ where the expected number of observations is $n^{1+\kappa}$ for any small constant $\kappa > 0$ and the $\tilde{O}$ notation simply hides polylog factors. Our result implies an infinity norm bound on the error rather than the looser typical mean squared error

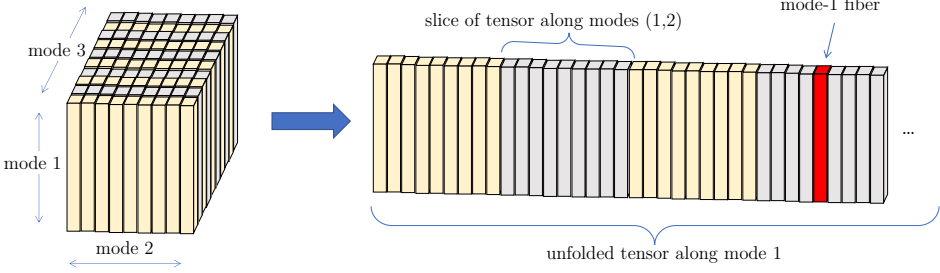


Fig. 1. Depicting an unfolding of a 3rd order tensor along mode 1, denoted as $T_{(1)}$. The columns of the resulting matrix are referred to as the mode-1 fibers of the tensor. The submatrices alternating in color are referred to as slices of the tensor along modes (1, 2).

bounds provided in the existing literature. We also provide experiments on both synthetic and real-world datasets that validate our theoretical insights.

2 PRELIMINARIES

Consider the task of estimating a t -order tensor T given observations $\{T^{obs}(\mathbf{i})\}_{\mathbf{i} \in \Omega}$ for some $\Omega \subset [n_1] \times [n_2] \times \dots \times [n_t]$ where $|\Omega|$ is significantly smaller than $n_1 n_2 \dots n_t$. Assume an additive noise model where $T^{obs}(\mathbf{i}) = T(\mathbf{i}) + E(\mathbf{i})$ for $\mathbf{i} \in \Omega$, and E is the noise matrix with independent mean zero entries. We assume a uniform Bernoulli sampling model where each entry is observed independently with probability p , i.e. $\mathbb{1}_{[\mathbf{i} \in \Omega]} \sim \text{Bernoulli}(p)$.

Let $T_{(\ell)}$ denote the unfolded tensor along the ℓ -th mode, which is a matrix of dimension $n_\ell \times \prod_{\ell' \in [t] \setminus \{\ell\}} n_{\ell'}$. We refer to columns of $T_{(\ell)}$ as mode- ℓ fibers of tensor T , which consists of vectors constructed by varying only the index in mode ℓ and fixing the indices along all other modes not equal to ℓ . We refer to slices of the tensor along modes (ℓ, ℓ') to be the matrix of entries resulting from varying the indices in modes ℓ and ℓ' and fixing the indices along all modes not equal to ℓ or ℓ' . Figure 1 visualizes one such unfolding of the tensor.

We must impose low dimensional structure on the underlying tensor to reasonably expect sparse tensor estimation to be feasible. Unlike in the matrix setting, there are multiple definitions of tensor rank. *CP-rank* is the minimum number of rank-1 tensors such that their sum is equal to the desired tensor. An *overcomplete* tensor is one for which the CP-rank is larger than the dimension n . The latent factors in the minimal rank-1 CP-decomposition may not be orthogonal. The *Tucker rank*, or *multilinear rank*, is a vector $(r_1, r_2, \dots, r_t)$ such that for each mode $\ell \in [t]$, r_ℓ is the dimension of the column space of $T_{(\ell)}$, the unfolded tensor along the ℓ -th mode to a $n_\ell \times \prod_{i \neq \ell} n_i$ matrix. The Tucker rank is also the minimal values of $(r_1, r_2, \dots, r_t)$ for which the tensor can be decomposed according to a multilinear multiplication of a core tensor $\Lambda \in \mathbb{R}^{r_1 \times r_2 \times \dots \times r_t}$ with latent factor matrices $Q_1 \dots Q_t$ for $Q_\ell \in \mathbb{R}^{n_\ell \times r_\ell}$, denoted as

$$T = (Q_1 \otimes \dots \otimes Q_t) \cdot (\Lambda) := \sum_{\mathbf{k} \in [r_1] \times [r_2] \times \dots \times [r_t]} \Lambda(\mathbf{k}) Q_1(\cdot, k_1) \otimes Q_2(\cdot, k_2) \dots \otimes Q_t(\cdot, k_t), \quad (1)$$

and depicted in Figure 2. The higher order SVD (HOSVD) specifies a unique Tucker decomposition in which the factor matrices $Q_1 \dots Q_t$ are orthonormal and correspond to the left singular vectors of the unfolded tensor along each mode. Furthermore, the slices of the core tensor Λ along each mode are mutually orthogonal with respect to entrywise multiplication, and the Frobenius norm of the slices of the core tensor are ordered decreasingly and correspond to the singular values of the matrix resulting from unfolding the tensor along each mode [15]. The direct relationship between

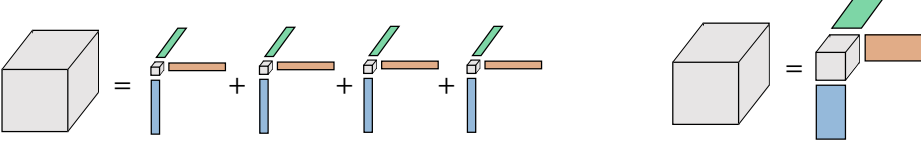


Fig. 2. (Left) The tensor CP-rank admits a decomposition corresponding to the sum of r rank-1 tensors. (Right) The Tucker rank or multilinear rank $(r_1, r_2, \dots, r_t)$ admits a decomposition corresponding to a multilinear multiplication of a core tensor of dimensions $(r_1, r_2, \dots, r_t)$ with latent factor matrices associated to each mode. The number of degrees of freedom in each model scales linearly with n via the latent factors.

the HOSVD and the SVD of each unfolded tensor provides a direct method to compute the HOSVD and thus the multilinear rank from complete observation of a tensor.

If the CP-rank is r , the Tucker-rank is bounded above by $(r, r, \dots, r)$ by constructing a superdiagonal core tensor. If the Tucker rank is $(r_1, r_2, \dots, r_t)$, the CP-rank is bounded by the number of nonzero entries in the core tensor, which is at most $r_1 r_2 \dots r_t / (\max_t r_t)$ [15]. While the latent factors of the HOSVD are orthogonal, the latent factors corresponding to the minimal CP-decomposition may not be orthogonal. For simplicity of presentation, we will first consider a limited setting where there exists a decomposition of the tensor into the sum of orthogonal rank-1 tensors. This is equivalent to enforcing that the core tensor Λ associated to the Tucker decomposition is superdiagonal, or equivalently enforcing that the latent factors in the minimal CP-decomposition are orthogonal. There does not always exist such an orthogonal CP-decomposition, however this class still includes all rank 1 tensors which encompasses the class of instances used to construct the hardness conjecture in [1]. Our results also extend beyond to general tensors as well, though the presentation is simpler in the orthogonal setting.

As some of our matrix and tensor variables will have subscripts, we will index entries in the matrix or tensor by arguments in parentheses, i.e. $Q_\ell(a, b)$ denoting the entry in matrix Q_ℓ associated to index (a, b) . We will use bold font $\mathbf{i} = (i_1, i_2, \dots, i_t)$ to denote the index vector of a tensor and i_ℓ to denote the ℓ -th coordinate of vector $\mathbf{i}$. For data vectors, we use parentheses notation to access entries in the vector. We use e_a to denote the standard basis vector with 1 at coordinate a and zero elsewhere, and we use $\mathbf{1}$ to denote the all ones vector. We denote the set $[n] = \{1, 2, \dots, n\}$. $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$. For $y \neq z \in [t]^2$, let $\mathcal{I}_{yz}(a, b) = \{\mathbf{i} \in [n_1] \times \dots \times [n_t] \text{ s.t. } i_y = a, i_z = b\}$ denotes the set of indices $\mathbf{i}$ such that the y -th coordinate is equal to a and the z -th coordinate is equal to b . Similarly define $\mathcal{J}_{yz}(a, b) = \{\mathbf{k} \in [r_1] \times \dots \times [r_t] \text{ s.t. } k_y = a, k_z = b\}$.

3 KEY INTUITION IN A SIMPLE SETTING

Consider a simple setting for a 3-order tensor that has low orthogonal CP-rank r , i.e.

$$T = \sum_{k=1}^r \lambda_k Q_1(\cdot, k) \otimes Q_2(\cdot, k) \otimes Q_3(\cdot, k), \quad (2)$$

where $\otimes$ denotes an outer product and the columns of the latent factor matrices $Q_1 \in \mathbb{R}^{n_1 \times r}$, $Q_2 \in \mathbb{R}^{n_2 \times r}$, and $Q_3 \in \mathbb{R}^{n_3 \times r}$ are orthonormal. This is equivalent to assuming a Tucker decomposition consisting of the latent factor matrices Q_1, Q_2 , and Q_3 , with a superdiagonal core tensor Λ having $(\lambda_1, \lambda_2, \dots, \lambda_r)$ on the superdiagonal. Suppose that you were given additional information that for each $k \in [r]$ and $\ell \in [3]$, $|\langle Q_\ell(\cdot, k), \frac{1}{n_\ell} \mathbf{1} \rangle| > \mu > 0$. Let us construct matrix M_{12}^{obs} by averaging

observed entries along the 3rd mode,

$$M_{12}^{obs}(a, b) = \frac{\sum_{i=1}^{n_3} T^{obs}(a, b, i) \mathbb{1}_{[(a,b,i) \in \Omega]}}{\sum_{i=1}^{n_3} \mathbb{1}_{[(a,b,i) \in \Omega]}}. \quad (3)$$

The expectation of M_{12}^{obs} with respect to the randomness in the sampling pattern and observation noise is a rank r matrix that has the same latent factors as the underlying tensor, i.e. conditioned on (a, b) being observed, $\mathbb{E}[M_{12}^{obs}(a, b)] = e_a^T Q_1 \tilde{\Lambda} Q_2^T e_b$ where $\tilde{\Lambda}$ is diagonal with $\tilde{\Lambda}_{kk} = \lambda_k \langle Q_3(\cdot, k), \frac{1}{n_3} \mathbf{1} \rangle$. The key observation is that the matrix $\mathbb{E}[M_{12}^{obs}(a, b)]$ and tensor T share the same latent factor matrices Q_1 and Q_2 . In Section 5.3, we show that this insight extends to the general context of a t -order tensor. Repeating this construction along other modes would result in matrices whose expected values share the latent factor matrix Q_3 with the tensor. A natural approach is to use these constructed matrices to estimate the latent factors of the original tensor along each mode. This insight is also used by [22] in the context of tensor factorization with a fully observed tensor.

Denote the sparsity pattern of matrix M_{12}^{obs} with $\tilde{\Omega}_{12}$, where $(a, b) \in \tilde{\Omega}_{12}$ iff there exists at least one $i \in [n_3]$ such that $(a, b, i) \in \Omega$, i.e. there is an observed tensor datapoint involving both coordinates a and b . Due to the Bernoulli sampling model, $\mathbb{1}_{[(a,b) \in \tilde{\Omega}_{12}]}$ is independent across all $(a, b) \in [n_1] \times [n_2]$, and $\mathbb{P}((a, b) \in \tilde{\Omega}_{12}) = 1 - (1 - p)^{n_3} =: \tilde{p} \approx pn_3$, where the approximation holds if $pn_3 = o(1)$. As a result, the density of observations in M_{12}^{obs} is $\tilde{p} = \Theta(\min(1, pn_3))$, which is significantly more dense than the original tensor dataset.

In this simple setting where $|\langle Q_\ell(\cdot, k), \frac{1}{n_3} \mathbf{1} \rangle| > \mu$, we have reduced the task of estimating the latent factor matrices of a sparsely observed tensor T to the easier task of estimating the latent factors of a not-as-sparsely observed matrix $\mathbb{E}[M_{12}^{obs}]$, where the data is also generated from a Bernoulli sampling model. As matrix estimation is very well understood, we can then apply methods from matrix estimation to learn the latent factors of $\mathbb{E}[M_{12}^{obs}]$. The non-zero singular values of $\mathbb{E}[M_{12}^{obs}]$ are $\lambda_k \langle Q_3(\cdot, k), \frac{1}{n_3} \mathbf{1} \rangle$, which has magnitude bounded below by $|\lambda_k| \mu$ by assumption from the additional information. If this inner product were equal to zero for any value of k , it would imply that the rank of $\mathbb{E}[M_{12}^{obs}]$ is strictly smaller than the r such that we would not be able to recover the full desired latent factor matrices for T by estimating $\mathbb{E}[M_{12}^{obs}]$.

The distribution of $M_{12}^{obs}(a, b)$ depends not only on the original additive noise model of the tensor, but also the sampling process and the latent factors Q_3 . For example the simplest rank-1 setting with exact observation of the tensor entries is depicted in Figure 3. The observation $M_{12}^{obs}(a, b)$ takes value $\lambda Q_1(a, 1) Q_2(b, 1) Z_{ab}$, where

$$Z_{ab} = \frac{\sum_{i=1}^{n_3} Q_3(i, 1) \mathbb{1}_{[(a,b,i) \in \Omega]}}{\sum_{i=1}^{n_3} \mathbb{1}_{[(a,b,i) \in \Omega]}}. \quad (4)$$

This highlights that the noise is in fact a multiplicative factor Z_{ab} , where $\mathbb{E}[Z_{ab}] = \langle Q_3(\cdot, 1), \frac{1}{n_3} \mathbf{1} \rangle$ and the independence in the original tensor observations also implies Z_{ab} are independent across indices (a, b) . Incoherence style assumptions on the latent factor guarantee that this multiplicative noise factor is not ill-behaved. While any matrix estimation algorithm could plausibly be used on M_{12}^{obs} , the theoretical results require analyses that are robust to more general noise models beyond the commonly assumed Gaussian additive noise model. When the data entries are bounded, we will use the approach in [4, 31, 32] as they provide statistical guarantees for general mean zero bounded noise.

Let's consider the impact on the sample complexity when the dimensions are equal, i.e. $n_1 = n_2 = \dots n_t = n$. In contrast to approaches that unfold the tensor to a matrix, our algorithm collapses the

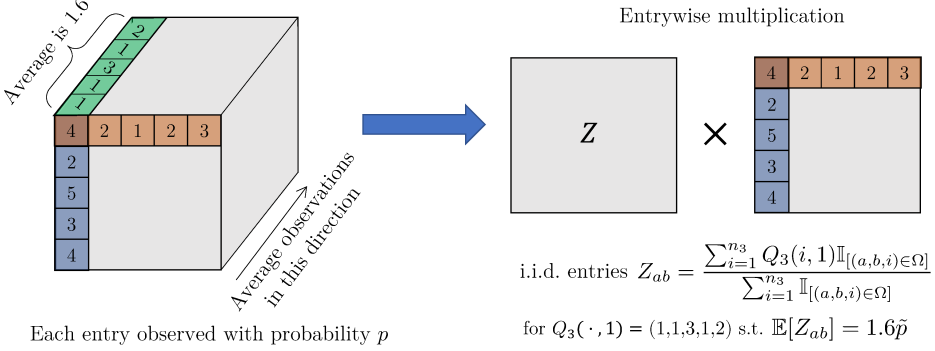


Fig. 3. For a rank 1 tensor with no additive noise, averaging out entries in the third mode results in a matrix whose expected value is a rank 1 matrix with the same latent factors as the original tensor.

tensor such that the constructed matrix has only $O(n^2)$ entries rather than $O(n^t)$ entries. Since matrix estimation algorithms require sample complexity scaling linearly in the maximum dimension of the matrix, this can be high for flattenings of the tensor to a matrix, in fact the maximum matrix dimension will always be at least $O(n^{t/2})$. By collapsing the tensor via averaging out other modes, the maximum matrix dimension of M_{12}^{obs} is only linear in n . As a result estimating $\mathbb{E}[M_{12}^{obs}]$ only requires roughly linear in n observations, which is significantly lower than the $O(n^{t/2})$ sample complexity common in tensor estimation.

To summarize, the key insight is that while there may not be sufficient observations to directly perform inference over the tensor T^{obs} due to high noise or sparsity, the constructed matrix M_{12}^{obs} can be significantly less noisy and sparse due to aggregating many datapoints. As low rank matrix completion algorithms only requires observing roughly linear number of entries in order to recover the underlying matrix, the goal is to show that we could recover the underlying tensor T when the number of observations is roughly linear in the dimension n , by applying matrix completion methods to appropriately constructed $n \times n$ matrices.

3.1 Sufficient Side Information

While we sketched the intuition for the simplest setting with low orthogonal CP-rank and the condition that $|\langle Q_\ell(\cdot, k), \frac{1}{n_\ell} \mathbf{1} \rangle| > \mu > 0$, this approach can be extended. The key property we needed was knowledge of some weight vector $W_\ell \in \mathbb{R}^{n_\ell}$ such that $|\langle Q_\ell(\cdot, k), \frac{1}{n_\ell} W_\ell \rangle| > \mu > 0$. Given such a weight vector, instead of computing M_{12}^{obs} according to a uniform averaging of observed entries as in (3), we simply compute a modified average using the given weight vector according to,

$$M_{12}^{obs}(a, b) = \frac{\sum_{i=1}^{n_3} T^{obs}(a, b, i) W_3(i) \mathbb{I}_{[(a,b,i) \in \Omega]}}{\sum_{i=1}^{n_3} \mathbb{I}_{[(a,b,i) \in \Omega]}}. \quad (5)$$

The expectation of M_{12}^{obs} would still share the same latent factor matrices as the tensor T , with $\mathbb{E}[M_{12}^{obs}(a, b)] = e_a^T Q_1 \tilde{\Lambda} Q_2^T e_b$ for $(a, b) \in \tilde{\Omega}$ where $\tilde{\Lambda}$ is diagonal with $\tilde{\Lambda}_{kk} = \lambda_k \langle Q_3(\cdot, k), \frac{1}{n_3} W_\ell \rangle$, which has magnitude bounded away from zero by assumption on W_ℓ . As a result, we could apply the same approach of estimating the latent factor matrices of the tensor using the more densely populated matrix $M_{12}^{obs}(a, b)$. In the same way that we will require incoherence style regularity conditions on Q_ℓ , under the weighted model, we additionally require regularity conditions on the weighted latent factor matrices $\text{diag}(W_\ell) Q_\ell$.

When does such side information in the form of these specified weight vectors exist? And when would it be reasonable to assume knowledge of such side information? The condition that $|\langle Q_\ell(\cdot, k), \frac{1}{n_\ell} W_\ell \rangle| > \mu > 0$ for all ℓ, k simply imposes that the properly scaled weight vector W_ℓ must not be nearly orthogonal to any of the latent factors $Q_\ell(\cdot, k)$. In comparison to other common models of side information, this is significantly weaker than assuming full knowledge of the column space of Q_ℓ . The requirement that W_ℓ is not directly aligned to the latent factors is mild, for example it would be satisfied by any linear combination of the latent factors with coefficients bounded away from zero, e.g. choosing a weight vector according to $\frac{1}{n_\ell} W_\ell = \sum_{k=1}^r \mu_\ell Q_\ell(\cdot, k)$ would satisfy $\langle Q_\ell(\cdot, k), \frac{1}{n_\ell} W_\ell \rangle = \mu_\ell$. If the latent factor matrices Q_ℓ satisfy incoherence-like properties, the vector W_ℓ would also satisfy similar regularity conditions. As a result, given observed features of the indices, it is relatively mild to assume that one can construct a vector that satisfies the desired properties for the side information.

When the tensor does not have low orthogonal CP-rank, i.e. the core tensor corresponding to the HOSVD is not superdiagonal, the equivalent condition for sufficient side information is slightly more involved. We discuss the formal conditions for general tensors in Section 5. Also, one may ask whether we could simply use the observed entries to construct a weight vector W_ℓ , in which case one would not need to be provided the weight vector in advance as auxiliary side information. Although this could be possible in special scenarios, a case study of the set of hard instances used in [1] shows that this is not possible in fullest generality, which we discuss in section 5.2.

4 ALGORITHM

Given sparse noisy observations of the tensor T^{obs} , and given side information in the form of a weight vector $W_\ell \in \mathbb{R}^{n_\ell}$ for each mode $\ell \in [t]$, the main approach is to transform the tensor estimation problem into a matrix problem where the latent factors of the matrix are the same as the latent factors of the tensor. Once we learn the latent factors of the tensor, we use the estimates to reconstruct the original tensor. While one could use different existing matrix algorithms for estimating the latent factor matrices given the constructed matrices, we choose to use a variant of the iterative collaborative filtering algorithm from [5] due to its ability to handle general bounded noise models.

Our algorithm uses a subroutine from the iterative collaborative filtering algorithm in [5] to compute distances $\hat{d}_y(a, b)$ that approximate $\|Q_y^T(e_a - e_b)\|_W$ for an appropriate W . Subsequently it uses those distance estimates to compute nearest neighbor estimates over the tensor dataset. For simpler notation, we assume that $n_1 = n_2 = \dots n_t = n$ and the density of observations $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$. The analysis can be modified to extend to tensors with uneven dimensions as long as the dimensions scale according to the same order. An additional benefit of using a nearest neighbor style algorithm is that the analysis leads to bounds on the maximum entrywise error of the final estimate as opposed to aggregate mean squared error bounds typical in the literature.

4.1 Formal Algorithm Statement

To facilitate cleaner analysis, we assume access to three fresh samples of the dataset, T_1^{obs} , T_2^{obs} , and T_3^{obs} associated to observation sets $\Omega_1, \Omega_2, \Omega_3$ respectively (this is not necessary and can be instead handled by sample splitting as illustrated in [5]). Each dataset is used for different part of the algorithm to remove correlation across computations in different steps. T_1^{obs} is used to construct M_{yz}^{obs} in (6); T_2^{obs} is used in the distance computation in (8), and T_3^{obs} will be used for the final estimates in (10).

Phase 1 (Estimating distances): For each mode $y \in [t]$, choose some $z \neq y$ and follow Steps 1-3 to estimate distances $\hat{d}_y(a, b)$ for each pair of $(a, b) \in [n_y]^2$ using the datasets T_1^{obs} and T_2^{obs} .

Step 1: Construct M_{yz}^{obs} and associated data graph. Construct M_{yz}^{obs} using data in T_1^{obs} according to

$$M_{yz}^{obs}(a, b) = \frac{\sum_{\mathbf{i} \in \mathcal{I}_{yz}(a, b)} T_1^{obs}(\mathbf{i}) \prod_{\ell \in [t] \setminus \{y, z\}} W_\ell(i_\ell)}{|\Omega \cap \mathcal{I}_{yz}(a, b)|}, \quad (6)$$

where recall that $\mathcal{I}_{yz}(a, b) = \{\mathbf{i} \in [n_1] \times \cdots \times [n_t] \text{ s.t. } i_y = a, i_z = b\}$, and $\mathbf{i} = (i_1, i_2, \dots, i_t)$. Let $\tilde{\Omega}_{yz}$ denote the index set of nonzero entries in M_{yz}^{obs} . Let $\mathcal{G}_{yz}$ denote the bipartite graph with vertex sets $[n_y]$ and $[n_z]$ and edge set $\mathcal{E}_{yz}$. The pair $(a, b) \in [n_y] \times [n_z]$ is an edge in $\mathcal{E}_{yz}$ iff $(a, b) \in \tilde{\Omega}_{yz}$.

Step 2: Construct BFS trees and associated statistics. For each $a \in [n_y]$, construct a breadth first search tree rooted at vertex a using edges in $\mathcal{G}_{yz}$. Let $\mathcal{S}_{a,s}$ denote the set of vertices in $\mathcal{G}_{yz}$ that are at distance s from root vertex a . Let $\mathcal{B}_{a,s}$ denote the set of vertices in $\mathcal{G}_{yz}$ that are at distance at most s from a , i.e. $\mathcal{B}_{a,s} = \cup_{h=0}^s \mathcal{S}_{a,h}$. Let $\text{path}(a, i)$ denote the set of edges along the shortest path from vertex a to i . Define neighborhood vectors

$$N_{a,s}(i) = \mathbb{1}_{[i \in \mathcal{S}_{a,s}]} \prod_{(u,v) \in \text{path}(a,i)} M_{yz}^{obs}(u, v). \quad (7)$$

Denote normalized neighborhood vectors as $\tilde{N}_{a,s} = N_{a,s}/|\mathcal{S}_{a,s}|$. Choose the depth $s = \lceil \ln(n)/\ln(pn^{t-1}) \rceil$.

Step 3: Distance computation. For each $a \neq b \in [n]^2$, compute distance estimates $\hat{d}_y(a, b)$ via

$$\hat{d}_y(a, b) = D(a, a) + D(b, b) - D(a, b) - D(b, a) \quad (8)$$

$$D(u, v) = \frac{1}{pn^{t-2}} \sum_{i,j} \tilde{N}_{u,s}(i) \tilde{N}_{v,s+1}(j) \sum_{\mathbf{h} \in \mathcal{I}(i,j)} T_2^{obs}(\mathbf{h}) \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(h_\ell) \quad (9)$$

where $\mathcal{I}(i, j) = \mathcal{I}_{yz}(i, j)$ for even values of s and $\mathcal{I}(i, j) = \mathcal{I}_{yz}(j, i)$ for odd values of s .

Step 4: Latent subspace computation (optional). While the distance estimates in $\hat{d}_y$ will be sufficient to compute the nearest neighbor estimator described below in Phase 2, if one desires to directly obtain an estimate for the latent subspaces, one could compute the singular value decomposition of the symmetric matrix corresponding to the distance estimates $\hat{d}_y$, and let $\hat{Q}_y$ denote the eigenvectors corresponding to the top r_y eigenvalues of the estimated distance matrix. $\hat{Q}_y$ serves as an r_y -dimensional approximation for the latent subspace along the y -th mode of the tensor.

Phase 2 (Nearest neighbor averaging): Given the distance estimates $\hat{d}_y(a, b)$ for all $y \in [t]$ and $(a, b) \in [n_y]^2$, estimate the tensor using nearest neighbor averaging,

$$\hat{T}(\mathbf{i}) = \frac{\sum_{\mathbf{i}' \in \Omega_3} T_3^{obs}(\mathbf{i}') K(\mathbf{i}, \mathbf{i}')}{\sum_{\mathbf{i}' \in \Omega_3} K(\mathbf{i}, \mathbf{i}')} \quad \text{for } K(\mathbf{i}, \mathbf{i}') = \prod_{\ell \in [t]} \mathbb{1}_{[d_\ell(i_\ell, i'_\ell) \leq \eta]}. \quad (10)$$

where we use a simple threshold kernel with parameter $\eta = \Theta(\max(n^{-\kappa/2}, n^{-2(\kappa+1)/(t+2)}))$.

Let $m = |\Omega|$ denote the number of observations in the tensor, and let $m' = \max_{y,z} |\tilde{\Omega}_{yz}|$ denote the maximum number of observations in the matrices that are constructed from averaging out $t-2$ modes of the tensor, where the maximum is taken over y, z . A naive upper bound on the computational complexity of the algorithm is $t(m + nm' + nm') + n'm$, where the terms are from constructing the matrix M_{yz}^{obs} , for constructing the BFS trees associated to $\mathcal{G}$, computing pairwise distances, and calculating the nearest neighbor estimator. One could improve the computational complexity of the algorithm by choosing only to compute pairwise distances between a subset of vertex pairs, and clustering the vertices. For an appropriate choice of number of clusters, this would achieve the same theoretical guarantees while having significantly reduced computation.

4.2 Variations of the Algorithm

We presented a specific algorithm above to illustrate how to use the given side information to simplify tensor completion. However, the basic approach of using the side information to reduce the tensor to a matrix along each mode can also lead to other natural algorithms arising from other matrix estimation algorithms, as presented below.

Phase 1 (Estimate latent subspaces): For each mode $y \in [t]$ and some $z \neq y$, construct the matrix M_{yz}^{obs} according to equation (6). Use any low rank matrix estimation algorithm to estimate $\mathbb{E} \left[M_{yz}^{obs} \right]$, denoting the output as $\hat{M}_{yz}$. Compute the singular value decomposition of $\hat{M}_{yz}$, and let $\hat{Q}_y$ denote the left singular vectors corresponding to the top r_y largest singular values of $\hat{M}_{yz}$. $\hat{Q}_y$ is as an r_y -dimensional approximation for the latent subspace along the y -th mode of the tensor.

Phase 2 (Least squares minimization): Using the matrices $\hat{Q}_y$ as an approximation for the latent factor matrices Q_y , the tensor can be approximated by a multilinear multiplication of the latent factor matrices $\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_t$ with a $r_1 \times r_2 \times \dots \times r_t$ core tensor. Compute $\hat{T}$ by solving the following unconstrained convex program minimizing the squared error

$$\text{minimize} \quad \sum_{\mathbf{i} \in \Omega} (T^{obs}(\mathbf{i}) - \hat{T}(\mathbf{i}))^2 \quad (11)$$

$$\text{s.t.} \quad \hat{T} = (\hat{Q}_1 \otimes \dots \otimes \hat{Q}_t) \cdot (\hat{\Lambda}) \quad (12)$$

$$\hat{\Lambda} \in \mathbb{R}^{r_1 \times r_2 \times \dots \times r_t} \quad (13)$$

Empirically, we find that both proposed variants of Phase 1 are not computationally costly (unless the matrix estimation subroutine chosen is itself costly), but the nearest neighbor averaging in Phase 2 can be slow. The least squares minimization variant of Phase 2 is both faster and seems to perform slightly better in practice, as it does not require tuning of hyperparameters, whereas nearest neighbor averaging is sensitive to the choice of the averaging threshold. The nearest neighbor averaging variant is still useful to present though as it facilitates an easier analysis for showing our significant gains in sample complexity.

5 THEORETICAL GUARANTEES

We will present theoretical guarantees for the algorithm presented in Section 4.1, which builds upon the iterative collaborative filtering style matrix estimation algorithm; as a result the model assumptions and analysis are similar to [5]. The first two are standard assumptions on uniform sampling and mean zero bounded observation noise.

ASSUMPTION 1 (SAMPLING MODEL). *The set of observations is drawn from a uniform Bernoulli(p) sampling model, i.e. each entry $\mathbf{i} \in [n_1] \times [n_2] \times \dots \times [n_t]$ is observed independently with probability p .*

ASSUMPTION 2 (OBSERVATION NOISE). *Each observation is perturbed by independent mean zero additive noise, i.e. $T^{obs}(\mathbf{i}) = T(\mathbf{i}) + E(\mathbf{i})$ where $E(\mathbf{i})$ are mean zero and independent across indices. Furthermore we assume boundedness such that $|T^{obs}(\mathbf{i})| \leq 1$.*

The boundedness on $T^{obs}(\mathbf{i})$ implies that the noise terms $E(\mathbf{i})$ are also bounded and also allow for heteroskedastic noise. The next assumption is in lieu of the standard incoherence assumption, which imposes regularity. We assume a latent variable model, in which the coordinates of each mode ℓ are associated to latent variables that are drawn from a population distribution, which for simplicity of notation is modeled as uniform over the unit interval. The population distribution can be easily extended to general distributions over a bounded set.

ASSUMPTION 3 (LATENT VARIABLE MODEL). For each mode $\ell \in [t]$, each coordinate $i_\ell \in [n_\ell]$ is associated to an i.i.d. sampled latent variable $x_\ell(i_\ell) \sim U[0, 1]$. Each mode is associated to a set of r bounded orthonormal functions $\{q_{\ell k}\}_{k \in [r]}$ with respect to $U[0, 1]$, i.e. $|q_{\ell k}(x)| \leq B$ and

$$\mathbb{E}[q_{\ell k}(x)q_{\ell h}(x)] = \int_0^1 q_{\ell k}(x)q_{\ell h}(x)dx = \mathbb{1}_{[k=h]}. \quad (14)$$

The ground truth tensor can be described by a bounded latent function f of the latent variables with finite spectrum r , where $|f(\mathbf{x})| \leq 1$ and f can be decomposed according to

$$T(\mathbf{i}) = f(x_1(i_1), x_2(i_2), \dots, x_t(i_t)) = \sum_{\mathbf{k} \in [r]^t} \Lambda(\mathbf{k}) q_{1k_1}(x_1(i_1)) q_{2k_2}(x_2(i_2)) \cdots q_{tk_t}(x_t(i_t)) \quad (15)$$

for some core tensor $\Lambda \in \mathbb{R}^{[r]^t}$.

The decomposition in (15) implies that the tensor T can be written according to a Tucker decomposition as in (1) with latent factor matrices $Q_\ell(i, k) = q_{\ell k}(x_\ell(i))$ and core tensor $\Lambda(\mathbf{k})$ having dimension $r \times r \times \cdots \times r$. This implies the multilinear rank or Tucker rank is bounded above by $(r, r, \dots, r)$. The multilinear rank of the tensor T will be given by the multilinear rank of the core tensor Λ . As formally stated, we assume the latent variables are sampled from $U[0, 1]$, but this can be relaxed to any general population distribution P_ℓ over a bounded set.

The latent variable model assumptions impose a distribution over the latent factors of the tensor such that in expectation over the randomness in the latent factors, they are orthogonal. The latent variables induce a distribution on the latent factors of the low rank decomposition via the functions $q_{\ell k}(x)$. For sufficiently large $n_1, n_2, \dots, n_t$, the distribution of the sampled latent variables will be close to that of the population distribution, such that the finite averages over the sampled latent variables will converge to the population mean over the latent variable distribution. As a result, $Q_1 \dots Q_t$ are random latent factor matrices that have orthogonal columns in expectation with respect to the latent variable distribution, and as $n_1, n_2, \dots, n_t \rightarrow \infty$ the columns of the sampled latent factor matrices of the tensor will be approximately orthogonal with high probability.

Along with the boundedness property, the generative model over the latent factors also guarantees incoherence style regularity conditions with high probability. Intuitively, incoherence attempts to formalize the concept that the signal is well spread amongst the coordinates and no single coordinate contains a critical component of the signal that is not exhibited by any other coordinate. When the latent factors are sampled from an underlying population distribution over a bounded set, then for sufficiently large n_ℓ , there will be other coordinates that exhibit similar latent variables, and thus carry similar information content. This assumption also helps to ensure that the ground truth tensor is not too empty. In particular, as the latent function is independent of the dimensions $n_1, \dots, n_t$ and sparsity p , even if the latent function f may take value zero for some subset of the latent feature space, it is still a constant measure with respect to the dimensions and sparsity parameters. As a result, the fraction of indices $\mathbf{i}$ for which $\mathbf{i} \in \Omega$ and $f(\mathbf{i}) \neq 0$ will scale as $\Theta(p)$.

ASSUMPTION 4 (LIPSCHITZNESS). The latent function $f(\mathbf{x})$ is L -Lipschitz with respect to the 1-norm over the latent vector $\mathbf{x} \in [0, 1]^t$.

The Lipschitz assumption is used only in the analysis of the final estimate computed from nearest neighbor averaging. This assumption can be replaced by conditions on the induced distribution over $Q_\ell(i, \cdot)$ guaranteeing that for any index $i \in [n_\ell]$, there exists sufficiently many indices $j \in [n_\ell]$ such that $\|Q_\ell^T(e_i - e_j)\|_2$ is small, such that averaging over similar indices in the nearest neighbor estimator will converge.

ASSUMPTION 5 (SIDE INFORMATION). *Side information is provided in the form of weight vectors $W_\ell \in [-1, 1]^{n_\ell}$ such that their values are consistent with associated weight functions $w_\ell : [0, 1] \rightarrow [-1, 1]$ evaluated on the latent variables, i.e. $W_\ell(i) = w_\ell(x_\ell(i))$. Furthermore, for each pair $y \neq z \in [t]^2$, the matrix $\hat{\Lambda}_{yz} \in \mathbb{R}^{r \times r}$ defined by*

$$\hat{\Lambda}(a, b) := \sum_{\mathbf{k} \in \mathcal{J}_{yz}(a, b)} \Lambda(\mathbf{k}) \prod_{\ell \in [t] \setminus \{y, z\}} \frac{1}{n} \sum_{i \in [n]} q_{\ell k}(x_\ell(i)) w_\ell(x_\ell(i)). \quad (16)$$

is well conditioned with high probability with respect to the latent variables, i.e. the condition number of the matrix $\tilde{\Lambda}$ is bounded by a constant. Recall that $\mathcal{J}_{yz}(a, b) = \{\mathbf{k} \in [r]^t \text{ s.t. } k_y = a, k_z = b\}$.

When the core tensor Λ is superdiagonal, i.e. only having nonzero entries in indices $\mathbf{k} = (k, k, \dots, k)$, then the assumption on side information is equivalent to assuming that for each $\ell \in [t]$ and $k \in [r]$, $\frac{1}{n} \sum_{i \in [n]} q_{\ell k}(x_\ell(i)) w_\ell(x_\ell(i))$ is bounded away from zero. Assumption 5 generalizes the intuition presented in Section 3 beyond low CP-orthogonal rank tensors as Λ does not need to be superdiagonal for this condition to be satisfied. However, Assumption 5 does require the multilinear rank of the core tensor Λ to be balanced and equal to $(r, r, \dots, r)$ in order for there to exist such a weight function such that $\tilde{\Lambda}_{yz}$ is full rank and well-conditioned for all (y, z) .

For simplicity of notation, we present our main theorem for $n_1 = n_2 = \dots = n_t = n$, however the result extends to uneven dimensions as well as long as they scale proportionally together.

THEOREM 1. *Assume the data generating model for the observation tensor T^{obs} satisfies Assumptions 1 to 4, and assume the side information $\{W_\ell\}_{\ell \in [t]}$ satisfies Assumption 5. Under sparse observations with $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$, with probability $1 - \frac{8(1+o(1))}{n}$, the max entrywise error of the estimate output by our algorithm is bounded above by*

$$\max_{\mathbf{i} \in [n]^t} |\hat{T}(\mathbf{i}) - T(\mathbf{i})| = O\left(\max\left(\frac{\log^{1/4}(n^2)}{n^{\min(\kappa, 1)/4}}, \frac{\log^{1/(t+2)}(n^{t+1})}{n^{(\kappa+1)/(t+2)}}\right)\right).$$

It follows then that the mean squared error is also bounded above by

$$\mathbb{E}\left[\frac{1}{n^t} \sum_{\mathbf{i} \in [n]^t} (\hat{T}(\mathbf{i}) - T(\mathbf{i}))^2\right] = O\left(\max\left(\frac{\log^{1/2}(n^2)}{n^{\min(\kappa, 1)/2}}, \frac{\log^{2/(t+2)}(n^{t+1})}{n^{2(\kappa+1)/(t+2)}}\right)\right).$$

Theorem 1 implies that given our model assumptions for ultra-sparse settings where the density $p = n^{-(t-1)+\kappa}$ for any arbitrarily small $\kappa > 0$, with high probability the max entrywise error of the estimate output by our algorithm goes to zero. Our result suggests that given appropriate side information, the estimation task is no harder for tensor estimation relative to matrix estimation, requiring only nearly linear sample complexity. This nearly linear sample complexity is a significant improvement from the best sample complexity of $O(n^{t/2})$ achieved by polynomial time algorithm, and is even better than the best achieved bound of $O(n^{3/2})$ of any statistical estimator. This form of side information only requires weak signal of the latent subspaces, and is significantly easier to satisfy than assuming full knowledge of the latent subspaces, as has been commonly assumed for previous empirical works studying tensor completion with side information.

The proof of the main theorem revolves around showing that the estimated distances $\hat{d}_y(a, b)$ concentrate around a function of the true distances with respect to the tensor latent factors, i.e. $\|AQ_y^T(e_a - e_b)\|_2^2$ for a well conditioned matrix A . Given bounds on the estimated distances, we only need to bound the performance of the nearest neighbor estimator, which follows from the independence of observation noise, the assumption that there are sufficiently many nearby datapoints, and a simple bias variance tradeoff.

5.1 Discussion

Our results assume a latent variable model and constant rank as it builds upon the algorithm and techniques from [4]. These assumptions replace the typical incoherence style conditions, as they guarantee that with high probability the latent factors are incoherent. This is reasonable for real-world settings with high dimensional data, as it essentially imposes regularity that the dimensions are associated to an underlying population distribution which satisfies regularity properties. The Lipschitz assumption is used to analyze the final nearest neighbor estimator, and is not necessary if we directly estimate the tensor from the latent factors.

Simulations suggest that the alternate variations of the algorithm using other matrix estimation algorithms also attain similar guarantees, which give us reason to believe that the theoretical guarantees also likely extend. However, the analysis of the matrix estimation algorithms would need to handle heteroskedastic noise (or just arbitrary bounded mean zero noise), as the variance across sampling of entries in the other $t - 2$ modes would depend on how widely the entries vary across the tensor. Additionally one would need guarantees on the row recovery of latent factors, which many of the previous work do not have. Recent work by [12] provides stronger control on recovery of latent factors, however it assumes a Gaussian noise model and would need to be extended to allow for arbitrary bounded noise that could result from aggregating entries sampled along other modes of the tensor.

The most restrictive assumption on the tensor is that the multilinear rank must be balanced, i.e. equal to $(r, r, \dots, r)$ for some constant value of r . When the multilinear rank is not evenly balanced, the proposed algorithm will not be able to fully recover information about the latent subspaces, since the dimension of the column space of $\tilde{\Lambda}_{yz}$ for any choice of weight vectors may be strictly lower than r_y , which is equal to the dimension of the column space of $T_{(y)}$, the unfolded tensor along the y -th mode. The limitation of our result follows from the specific construction of the matrix M_{yz}^{obs} by computing a weighted averaging along all other modes of the tensor except y and z . In particular, the i -th column of M_{yz}^{obs} consists of a weighted average of mode- y fibers of the tensor, but specifically restricted to fibers corresponding to the i -th column of the slices of the tensor along modes (y, z) , weighted according to the same weight vector. As such, the dimension of the column space of the expected matrix $\mathbb{E} \left[M_{yz}^{obs} \right]$ is limited by the dimension of the column space of unfolding of the tensor along mode- z , which could be smaller if $r_z < r_y$. A possible extension of our idea that we leave for future exploration is to construct a different type of matrix M_y^{obs} which consist of columns derived from weighted combinations of the mode- y fibers, not limited to aligning fibers along slices of the tensor. The goal would be to construct M_y^{obs} such that the column space of expected matrix $\mathbb{E} \left[M_y^{obs} \right]$ is equal to the column space of $T_{(y)}$. This seems plausible as the mode- y fibers span the desired subspace. The weighted combinations should combine a significant number of the mode- y fibers in order to guarantee that even when the dataset is only sparsely sampled, the density of observations in M_y^{obs} is sufficiently large.

5.2 In the Absence of Side Information

The requirements of the side information for our result are weak and minimal in contrast to the assumptions made by other works in the literature, which either assume knowledge of the latent subspace [2, 6, 10, 29, 43], or similarity or kernel matrices that express relationship amongst the coordinates [23, 28]. We only require knowledge of a weight vector for which taking a weighted average along the specified mode does not zero out any component of the signal, which is far from requiring knowledge of the latent subspaces. One may ask whether it could be possible to construct the side information perhaps by sampling random weight vectors, or computing it as a function of

the data itself. If this were possible, it would suggest an algorithm achieving nearly linear sample complexity even without side information. In general this is not possible without further conditions restricting the class of possible tensors. For example, consider the class of rank-1 tensors used to argue the hardness result from [1]. The latent factor vector $\theta \in \{-1, +1\}^n$ consists of randomly sampled entries where $\theta(i)$ for $i \in [n]$ is -1 with probability $1/2$ and $+1$ with probability $1/2$. As a result, for any choice of weight vector $W \in \mathbb{R}^n$ which is not too sparse, the expected value of $\langle \theta, W \rangle$ is zero, and furthermore with high probability, the typical vector θ will also result in a small inner product with the weight vector due to the symmetry in the generating distribution. As a result, the constructed matrices M_{yz}^{obs} would have an expected value of zero, clearly not preserving the latent subspace. Any class of tensors with such symmetry would also exhibit similar properties; for example if we sampled the latent factor matrices by normalizing random matrices where each entry is sampled from an independent mean-zero Gaussian, with high probability the latent factors would be orthogonal to any fixed or randomly sampled weight vector. However, one could argue that such perfect symmetry is in fact not common in real-world datasets. In section 6, we benchmark our algorithm on both synthetic and real-world datasets, and we find that this assumption on side information is in fact satisfied for the real-world datasets derived from traffic and MRI measurements. We conjecture that in fact many real-world datasets are not perfectly symmetric, such that the naive all ones weight vector would perform well; this only requires that the sums of the latent factors are not too close to zero.

The ability to collect a limited amount of active samples could alternately replace the requirement for side information. If for each mode $\ell \in [t]$ we actively sampled an entire mode- ℓ fiber at random, then the sampled vector could be used as the weight vector W_ℓ . The mode- ℓ fiber lies in the desired column space by definition, and by sampling a fiber at random, it is likely that the sampled vector would satisfy the required conditions. This suggests that given a budget of tn active samples, we could produce a consistent estimate of the full tensor with nearly linear sample complexity, in contrast to the $n^{t/2}$ sample complexity without the active samples. This is consistent to the result in [40], which states that with an active sampling scheme, one can estimate the full tensor with $O(rtn)$ samples by directly estimating the latent subspaces from the actively sampled datapoints. Our result shows that limited ability to collect samples actively, can still be immensely beneficial even if the active samples are not sufficient themselves to fully recover the tensor.

5.3 Proof Sketch

While we defer the formal proof to the appendix, we give the rough intuition and sketch of the proof here. The main property that the approach hinges on is that the column space of $\mathbb{E} \left[M_{yz}^{obs} \right]$ is the same as the column space of $T_{(y)}$, and furthermore the matrix M_{yz}^{obs} resulting from collapsing the tensor satisfies the desired regularity conditions in order to use a similar analysis as [5].

First, we show that the expected value of the constructed matrix, with respect to the additive noise terms and the sampling distribution conditioned on the latent variables, indeed shares the same column space as the original tensor.

$$\mathbb{E} \left[M_{yz}^{obs}(a, b) \mid (a, b) \in \tilde{\Omega}, \{x_\ell(i)\} \right] \quad (17)$$

$$= \mathbb{E} \left[\frac{\sum_{i \in I_{yz}(a, b)} T_1^{obs}(i) \prod_{\ell \in [t] \setminus \{y, z\}} W_\ell(i_\ell)}{|\Omega \cap I_{yz}(a, b)|} \mid (a, b) \in \tilde{\Omega}, \{x_\ell(i)\} \right] \quad (18)$$

$$= \sum_{i \in I_{yz}(a, b)} \mathbb{E} \left[T(i) \prod_{\ell \in [t] \setminus \{y, z\}} W_\ell(i_\ell) \mid \{x_\ell(i)\} \right] \mathbb{E} \left[\frac{\mathbb{1}_{[i \in \Omega]}}{|\Omega \cap I_{yz}(a, b)|} \mid (a, b) \in \tilde{\Omega} \right] \quad (19)$$

$$= \sum_{i \in I_{yz}(a,b)} \sum_{\mathbf{k} \in [r]^t} \Lambda(\mathbf{k}) q_{y\mathbf{k}_y}(x_y(a)) q_{z\mathbf{k}_z}(x_z(b)) \prod_{\ell \in [t] \setminus \{y,z\}} q_{\ell\mathbf{k}_\ell}(x_\ell(i_\ell)) w_\ell(x_\ell(i_\ell)) \frac{1}{n^{t-2}} \quad (20)$$

$$= \sum_{j,h \in [r]^2} q_{yj}(x_y(a)) q_{zh}(x_z(b)) \sum_{\mathbf{k} \in \mathcal{I}_{yz}(j,h)} \Lambda(\mathbf{k}) \prod_{\ell \in [t] \setminus \{y,z\}} \frac{1}{n} \sum_{i=1}^n q_{\ell\mathbf{k}_\ell}(x_\ell(i)) w_\ell(x_\ell(i)) \quad (21)$$

$$= e_a^T Q_y \hat{\Lambda} Q_z^T e_b \quad (22)$$

for matrix $\hat{\Lambda}$ as defined in (16). If we take expectation with respect to the latent variables, then it follows that $\hat{\Lambda}$ concenetrates around $\tilde{\Lambda}_{yz}$, which is full rank by Assumption 5. Let the SVD of $\hat{\Lambda}$ be denoted $\hat{U} \hat{\Sigma} \hat{V}^T$, where the diagonal entries of $\hat{\Sigma}$ are denoted as $\hat{\sigma}_1 \dots \hat{\sigma}_r$. Let $\tilde{Q}_y = Q_y \hat{U}$ and $\tilde{Q}_z = Q_z \hat{V}$ with the associated latent functions $\tilde{q}_{yk}, \tilde{q}_{zk}$ defined similarly. It follows then that M_{yz}^{obs} is modeled by a latent variable model where the latent function is low rank and L -Lipshitz.

$$\mathbb{E} \left[M_{yz}^{obs}(a, b) \mid (a, b) \in \tilde{\Omega} \right] = e_a^T \tilde{Q}_y \tilde{\Sigma} \tilde{Q}_z^T e_b =: \tilde{f}(x_y(a), x_z(b)). \quad (23)$$

We can verify that the associated latent functions also exhibit orthonormality, e.g.

$$\int_0^1 \tilde{q}_{yk}(x) \tilde{q}_{yh}(x) dx = \int_0^1 \left(\sum_{i \in [r]} q_{yi}(x) \hat{U}(i, k) \right) \left(\sum_{j \in [r]} q_{yj}(x) \hat{U}(j, h) \right) dx \quad (24)$$

$$= \sum_{i,j \in [r]^2} \hat{U}(i, k) \hat{U}(j, h) \int_0^1 q_{yi}(x) q_{yj}(x) dx = \mathbb{1}_{[k=h]}. \quad (25)$$

Let $\tilde{B}$ be an upper bound on the latent functions $\tilde{q}_{\ell k}$ such that $|\tilde{q}_{\ell k}(x)| \leq \tilde{B}$. Due to boundedness of the functions $q_{\ell k}$, $\tilde{B}$ is at most Br . As a result, the latent function associated to $\tilde{f}(x_y(a), x_z(b))$ is rank r with the decomposition

$$\tilde{f}(x, x') = \sum_{k=1}^r \hat{\sigma}_k \tilde{q}_{yk}(x) \tilde{q}_{yk}(x'). \quad (26)$$

Furthermore we can verify that $\tilde{f}$ is also L -Lipschitz, inheriting the property from the latent function f . We use $f(x_y, x_z, \mathbf{x}_{-yz}(\mathbf{i}))$ to denote the latent function f evaluated at the vector having values x_y and x_z in the y and z coordinates, and having values $x_\ell(i)$ for all other coordinates $\ell \in [t] \setminus \{y, z\}$.

$$\begin{aligned} |\tilde{f}(x_y, x_z) - \tilde{f}(x'_y, x'_z)| &= \left| \frac{1}{n^{t-2}} \sum_{\mathbf{i} \in [n]^{t-2}} (f(x_y, x_z, \mathbf{x}_{-yz}(\mathbf{i})) - f(x'_y, x'_z, \mathbf{x}_{-yz}(\mathbf{i}))) \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(i_\ell) \right| \\ &\leq L|x_y - x'_y| + L|x_z - x'_z| \end{aligned} \quad (27)$$

where the last inequality follows from Lipschitzness of f as well as the boundedness assumption on $|W_\ell(i_\ell)| \leq 1$. This setup nearly satisfies the same data generating model that is used in [5]. A minor difference is that our model is asymmetric, however this leads to very little change in the analysis. The key difference in our model is that the distribution over $M_{yz}^{obs}(a, b)$ is more involved as it arises from the averaging of values over the other modes of the tensor. Furthermore we consider a wider range of densities $\tilde{p}$, from as sparse as $O(n^{-1+\kappa})$ to as dense as constant.

The proof relies on Lemma 1, stated below, which shows that the estimated distances that our algorithm produces will approximate a function of the true distances with respect to the tensor latent factors. The first step is to argue that the sparsity of the observed local neighborhood graph grows sufficiently quickly, which follows from standard arguments for the properties of an Erdos-Renyi graph that results from the Bernoulli sampling model. The second step is a careful martingale

concentration argument to show that $e_k^T \tilde{Q}_y \tilde{N}_{a,j} \approx e_k^T \hat{\Sigma}^j \tilde{Q}_y e_a$. The third step is to use show that $D(a, a')$ concentrates around $\tilde{N}_{a,s}^T \tilde{Q}_y^T \hat{\Sigma} \tilde{Q}_z \tilde{N}_{a',s+1}$. Standard applications of concentration inequalities yield results that are too loose when p is very sparse, i.e. $p = n^{-(t-1)+\kappa}$ for $\kappa \in (0, 1)$. As a result we define a “truncated” modification of $D(a, a')$, which can be shown to concentrate well via a standard application of Bernstein’s inequality, and we use the sparsity condition to argue that the truncated modification of $D(a, a')$ is in fact equal to $D(a, a')$ with high probability.

LEMMA 1. Assume $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$. Let $\tilde{p} = 1 - (1 - p)^{n^{t-2}}$. With probability $1 - \frac{6(1+o(1))}{tn}$, $\hat{d}_y(a, a')$ computed from (8) satisfies

$$\max_{a, a' \in [n]^2} |\hat{d}_y(a, a') - \|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2| \leq \frac{5rB\hat{\sigma}_{\max}^{s+1}(1 + \hat{\sigma}_{\max}^s) \log^{1/2}(trn^2)(1 + o(1))}{(n\tilde{p})^{1/2}}.$$

Next Lemma 2 uses the Lipschitz and latent variable model assumptions to argue that there are sufficiently many nearest neighbors with respect to the distances defined by $\|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2$.

LEMMA 2. For any $i \in [n]$, $\ell \in [t]$, $\eta' > 0$

$$\mathbb{P} \left(\sum_{i' \neq i \in [n]} \mathbb{1}[\|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2 \leq \eta'] \leq (1 - \delta)(n - 1) \frac{\sqrt{\eta'}}{\hat{\sigma}_{\max}^s L} \right) \leq \exp \left(- \frac{\delta^2(n - 1)\sqrt{\eta'}}{2\hat{\sigma}_{\max}^s L} \right).$$

The final steps of the proof of the main theorem follow from a straightforward analysis of the nearest neighbor averaging estimator given the estimated distances.

6 EXPERIMENTS

We provide a sequence of experiments ranging from synthetic to real-world datasets in order to understand the performance of our proposed algorithm as compared against state-of-the-art tensor completion algorithms. We benchmark the algorithms in the following four datasets.

Synthetic: We generate Gaussian matrices $U, V, W \in \mathbb{R}^{100 \times 10}$, with entries that are independently sampled from $N(1, 1)$. We sample a diagonal core tensor $S \in [0, 1]^{10 \times 10 \times 10}$, where each of the 10 diagonal entries is sampled from $U[0, 1]$. Let $Z = (U \otimes V \otimes W) \cdot (S)$ be the multilinear multiplication of the core tensor S with the latent factor matrices U, V, W . We then construct the ground truth tensor by normalizing Z so that $|T(i, j, k)| \leq 1$, according to $T = Z / \max_{i,j,k} |Z(i, j, k)|$. The ground truth tensor has rank 10 and dimensions $100 \times 100 \times 100$. For each observed location $(i, j, k) \in \Omega$, we add zero mean Gaussian observation noise with variance σ^2 , for $\sigma = 0.1$.

MRI: We construct a dataset in which the tensor T is a scaled version of a volumetric MRI brain-scan. We used a volumetric MRI brain-scan dataset from the MIRIAD dataset, which contains brain scans of Alzheimer’s sufferers and healthy elderly people [26]. We used a single MRI scan from the dataset, which is represented as a $256 \times 256 \times 124$ tensor. We normalized the data such that the absolute value is bounded by 1. We do not add additional noise to the measurements, but study the performance under sparse sampling.

Traffic: We use the urban traffic speed dataset of Guangzhou, China from [11]. This is a dataset which consists of speed measurements from 214 anonymous road segments (consisting of urban expressways and arterials) from Guangzhou, China. The measurements were collected at 10-minute intervals from August 1, 2016 to September 30, 2016. The dataset is represented as a $214 \times 61 \times 144$ tensor, where the modes correspond to road segment, day, and time window. We normalized the data such that the absolute value is bounded by 1. We do not add additional noise to the measurements, but study the performance under sparse sampling.

Random 3XOR: We generate an instance of tensor completion which embeds a 3XOR formula, as described in [1]. We randomly sample a latent vector $a \in \{-1, +1\}^n$, where a_i independently takes values -1 or $+1$ each with probability $1/2$. For each observed index $(i, j, k) \in \Omega$,

$$T^{obs}(i, j, k) = \begin{cases} a_i a_j a_k & \text{with prob } 15/16, \\ -a_i a_j a_k & \text{with prob } 1/16. \end{cases}$$

The expected tensor is rank 1, given by $T = \frac{7}{8} a \otimes a \otimes a$, such that $\mathbb{E}[T^{obs}(\mathbf{i})] = T(\mathbf{i})$ for all $\mathbf{i} \in \Omega$.

We will use the naïve all ones weight vector for the “side information”. While this does not inherently encode extra information, we will see that the algorithm performs well for the Synthetic, MRI, and Traffic datasets. To verify the required condition on the weight vector, we took each of the ground truth tensors T , computed the true latent factor matrices, and computed the inner product between the column in the latent factor matrices and the all ones vector. For the Synthetic, MRI, and Traffic datasets, the minimum inner product across all modes and latent factors is 0.135, 0.0149, and 0.000992, respectively. This supports our conjecture that real-world datasets are not perfectly symmetric, such that a naïve choice of weight vector may indeed suffice. In contrast, the inner product between the all ones vector and the latent factor in the 3XOR setting is as small as 9.71×10^{-17} . Consistent with the intuition from our analysis, our algorithm performs poorly for the random 3XOR dataset, as the latent factor is orthogonal to the weight vector.

6.1 Implementation Details of Our Proposed Algorithm

The key idea of our proposed algorithm is to reduce the tensor latent subspace estimation task to the simpler matrix latent subspace estimation task over a small matrix. We include implementations of three variations of our algorithm arising from different choices for Phases 1 and 2 of the algorithm.

- **ICF_NN** denotes the original algorithm as presented in Section 4.1, which consists of estimating distances in Phase 1 followed by nearest neighbor averaging in Phase 2.
- **ICF_LS** denotes the algorithm which uses Phase 1 from Section 4.1, which estimates latent subspaces via the estimated distances as described in Step 4, followed by least squares minimization for Phase 2 as presented in Section 4.2.
- **ALS_LS** denotes the algorithm presented in Section 4.2, using alternating least squares [39] for matrix estimation in Phase 1, followed by least squares minimization in Phase 2.

The depth parameter used in steps 2 and 3 of estimating distances in Phase 1 is chosen to be $s = 1$. The threshold for the nearest neighbor averaging in Phase 2 is chosen to be the 10th percentile of the estimated distances. We found that the algorithm performed well even without optimizing the hyperparameters; naturally the least squares variant (which has no hyperparameters) performed better than the nearest neighbor variant with the above fixed choice of hyperparameters. The rank hyperparameter for least squares minimization in Phase 2 is set to 10 along each mode for the Synthetic, MRI, and Traffic datasets, and set to 1 along each mode for the random 3XOR dataset.

6.2 Other Algorithms for Comparison

As a simple baseline, we compare with the naïve average, which simply fills in all missing entries with the average of all observed datapoints. When the data is extremely sparse, the benchmarked algorithms start to degrade significantly such that they perform worse than the naïve average. In addition, we compare against the following state of the art tensor completion algorithms. As all of these algorithms are iterative, the stopping criteria is met after a certain number of iterations (specified below) or when the difference between the iterates falls below a tolerance of 10^{-5} .

TenALS: Tensor alternating least squares (TenALS) uses an alternating minimization based method to solve a non-convex least square problem, initializing with a tensor power method [20]. The authors prove that under incoherence assumptions, TenALS can provably recover a three-mode $n \times n \times n$ dimensional rank- r tensor exactly from $O(n^{3/2}r^5 \log^4 n)$ randomly sampled entries [20]. We set the rank hyperparameter of TenALS to 10 for the Synthetic, MRI, and Traffic datasets, and we set it to 1 for the random 3XOR dataset. We use the default of 10 initializations and 50 iterations.

LRTC: Low rank tensor completion (LRTC) refers to a class of algorithms minimizing a variant of the tensor nuclear norm, which is a weighted average of the nuclear norms of all matrices unfolded along each mode [25]. The authors do not provide statistical guarantees for their algorithms. We benchmark against SiLRTC, which uses block coordinate descent, and HaLRTC, which uses alternating direction method of multipliers (ADMM). For SiLRTC, we set the hyperparameters as: $\alpha = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\beta = (1, 1, 1)$. Due to slow convergence, we set the number of maximum iterations to be 1000. For HaLRTC, we set the weights as $\alpha = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. We adaptively increase the Lagrangian multiplier, setting the initial value to be 10^{-2} , and increasing the multiplier by a factor of 1.1 in each iteration. The upper bound for the multiplier is set to be 10^{10} , and the number of maximum iterations is set to be 500.

TNN: [42] proposed a tensor nuclear norm (TNN) minimization algorithm which uses a definition of tensor nuclear norm constructed from the tensor-Singular Value Decomposition (t-SVD) proposed in [21]. Under incoherence conditions, TNN can provably recover a three-mode $n \times n \times n$ dimensional rank- r tensor exactly from $O(n^2r \log n)$ randomly sampled entries [41]. To minimize the tensor nuclear norm penalized objective function, they employ the general framework of ADMM. We set the initial value of the Lagrange multiplier to be 10^{-2} , and increase it gradually by a factor of 1.1 until it reaches 10^{10} . The number of maximum iterations is set to be 500.

6.3 Empirical Results

In each of these experiments, we specifically pay attention to the performance at extreme levels of data sparsity, close to the threshold where the algorithms “breaks”, i.e. performs worse than the naïve average baseline. As a result, we benchmark on values of $p \in [0, 0.08]$. We evaluate the algorithm’s performance with respect to the normalized mean squared error (MSE), defined as

$$\text{Normalized MSE} := \frac{\sum_{i \in [n_1]} \sum_{j \in [n_2]} \sum_{k \in [n_3]} \left(\hat{T}(i, j, k) - T(i, j, k) \right)^2}{\sum_{i \in [n_1]} \sum_{j \in [n_2]} \sum_{k \in [n_3]} T(i, j, k)^2}$$

Figure 4 plots the normalized MSE achieved by the output of each algorithm as a function of the data sparsity p . The scale of the x-axis for each plot are set differently in order to highlight the results at the sparsest regimes near the lower threshold of performance. The scale of the y-axis for each plot are set to highlight the performance relative to the naïve average baseline, which is depicted by the red dashed line. Let us first focus on the results from the Synthetic, MRI, and Traffic datasets. Firstly, we notice that ICF_LS consistently performs the best in the sparsest regimes relative to any other algorithm, followed by ALS_LS and ICF_NN. This suggests that our proposed algorithm in fact does do well in addressing data-poor settings. TenALS and HaLRTC perform very poorly in these sparse regimes, at times having MSE that is orders of magnitude higher than the average baseline (and thus not depicted in Figure 4). TNN and SiLRTC give reasonable performance, though it degrades more quickly than our proposed algorithms as the data becomes sparser. These experiments validate that our approach of reducing the tensor to a significantly smaller matrix to estimate the latent subspaces in fact does allow the algorithm to perform well at extreme levels of sparsity, even when the “side information” weight vector is naïvely chosen to be the all ones vector.

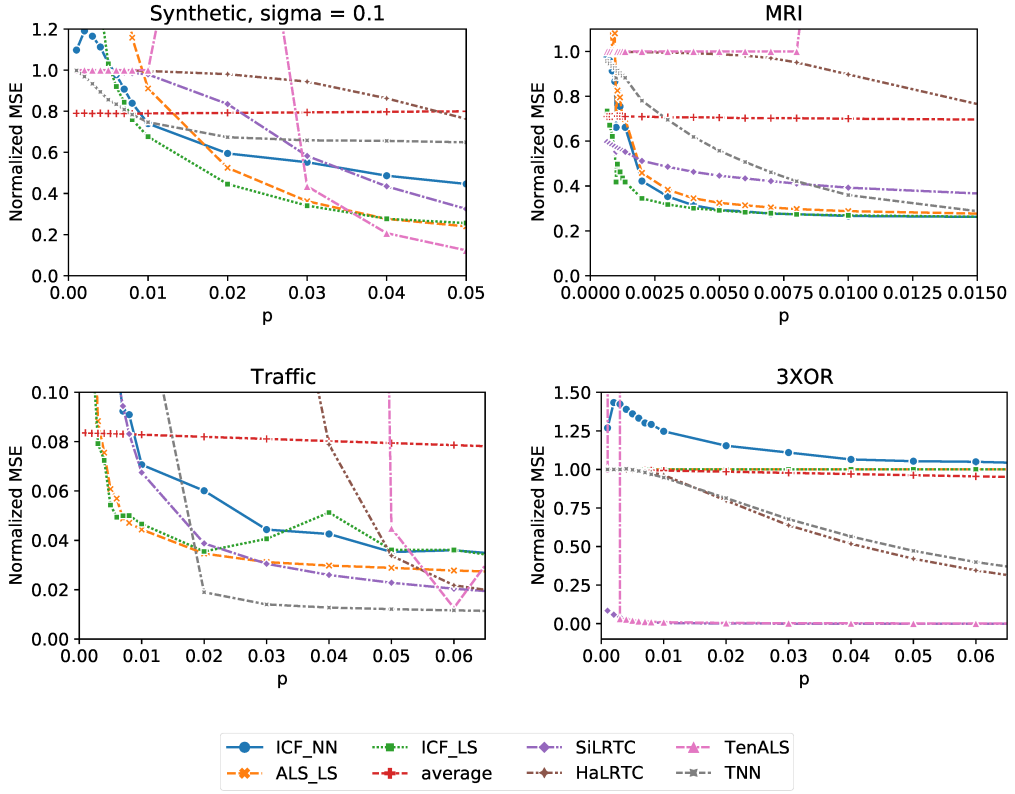


Fig. 4. These plots show the normalized mean squared error attained by different tensor completion algorithms on a synthetic dataset, an MRI dataset, a traffic dataset, and a random 3XOR dataset. The scaling of the y and x axis are chosen to show the performance relative to the naïve average baseline on the extremely sparse settings near the limits of the algorithms’ performance.

These examples also indicate that the side information requirements are indeed weak, and may not truly require “additional information” as long as the underlying model is not too symmetric.

In contrast, the random 3XOR dataset shows very different results. Our algorithms perform no better than the naïve average baseline, while TNN and HaLRTC show reasonably smooth degrading performance, and SiLRTC and TenALS perform extremely well even at low levels of sparsity. In fact, this example is chosen to illustrate when our algorithm will completely fail. The ground truth tensor is only rank-1, and the size of the tensor is not too large, i.e. $100 \times 100 \times 100$, so any brute force algorithm could perform very well. However, the entries in the latent factor are sampled independently from $\{-1, +1\}$, such that the latent factor is essentially orthogonal to the all ones vector. As our algorithm collapses the tensor along the third mode by summing according to the weight vector, this step unfortunately zeros out all the signal, resulting in a matrix that is essentially all zeros. As a result, our algorithms perform horribly poor on the random 3XOR dataset at all levels of sparsity p . Furthermore, it follows that without having access to secret knowledge of the latent factors, any attempt to construct a suitable weight vector for our algorithm will fail, as the symmetry in the distribution that the latent factors are sampled from will result in the latent factor being orthogonal in expectation to any chosen weight vector.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the NSF under grants CCF-1948256 and CNS-1955997. Christina Lee Yu is supported by an Intel Rising Stars Faculty Award. Data used in the preparation of this article were obtained from the Urban Traffic Speed Dataset of Guangzhou, China published by Xinyu Chen, Yixian Chen and Zhaocheng He, and the MIRIAD database. The MIRIAD investigators did not participate in analysis or writing of this report. The MIRIAD dataset is made available through the support of the UK Alzheimer's Society (Grant RF116). The original data collection was funded through an unrestricted educational grant from GlaxoSmithKline (Grant 6GKC).

REFERENCES

- [1] Boaz Barak and Ankur Moitra. 2016. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*.
- [2] Dimitris Bertsimas and Colin Pawłowski. [n.d.]. Tensor Completion with Noisy Side Information. ([n.d.]).
- [3] Srinadh Bhojanapalli and Sujay Sanghavi. 2015. A new sampling technique for tensors. *arXiv preprint arXiv:1502.05023* (2015).
- [4] Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. 2017. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Advances in Neural Information Processing Systems*. 4715–4726.
- [5] Christian Borgs, Jennifer T Chayes, Devavrat Shah, and Christina Lee Yu. 2021. Iterative collaborative filtering for sparse matrix estimation. *Operations Research* (2021).
- [6] Stanislav Budzinskiy and Nikolai Zamarashkin. 2020. Note: low-rank tensor train completion with side information based on Riemannian optimization. *arXiv preprint arXiv:2006.12798* (2020).
- [7] M Burkina, I Nazarov, M Panov, G Fedonin, and B Shirokikh. 2021. Inductive Matrix Completion with Feature Selection. *Computational Mathematics and Mathematical Physics* 61, 5 (2021), 719–732.
- [8] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. 2021. Nonconvex low-rank tensor completion from noisy data. *Operations Research* (2021).
- [9] Jian-Feng Cai, Lizhang Miao, Yang Wang, and Yin Xian. 2020. Provable Near-Optimal Low-Multilinear-Rank Tensor Recovery. *arXiv preprint arXiv:2007.08904* (2020).
- [10] Huiyuan Chen and Jing Li. 2019. Collective tensor completion with multiple heterogeneous side information. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 731–740.
- [11] Xinyu Chen, Yixian Chen, and Zhaocheng He. 2018. *Urban Traffic Speed Dataset of Guangzhou, China*. <https://doi.org/10.5281/zenodo.1205229>
- [12] Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. 2019. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* 116, 46 (2019), 22931–22937.
- [13] Kai-Yang Chiang, Inderjit S Dhillon, and Cho-Jui Hsieh. 2018. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *The Journal of Machine Learning Research* 19, 1 (2018), 3005–3039.
- [14] Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. 2015. Matrix Completion with Noisy Side Information.. In *NIPS*. 3447–3455.
- [15] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* 21, 4 (2000), 1253–1278.
- [16] Shmuel Friedland and Lek-Heng Lim. 2014. Nuclear norm of higher-order tensors. *arXiv preprint arXiv:1410.6072* (2014).
- [17] Silvia Gandy, Benjamin Recht, and Isao Yamada. 2011. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* 27, 2 (2011), 025010.
- [18] Mohsen Ghassemi, Anand Sarwate, and Naveen Goela. 2018. Global optimality in inductive matrix completion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2226–2230.
- [19] Prateek Jain and Inderjit S Dhillon. 2013. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626* (2013).
- [20] Prateek Jain and Sewoong Oh. 2014. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*. 1431–1439.
- [21] Misha Elena Kilmer, Karen S. Braman, Ning Hao, and Randy C. Hoover. 2013. Third-Order Tensors as Operators on Matrices: A Theoretical and Computational Framework with Applications in Imaging. *SIAM J. Matrix Anal. Appl.* 34 (2013), 148–172.
- [22] Tamara G Kolda. 2015. Symmetric orthogonal tensor decomposition is trivial. *arXiv preprint arXiv:1503.01375* (2015).
- [23] Hemank Lamba, Vaishnavh Nagarajan, Kijung Shin, and Naji Shajarisaless. 2016. Incorporating side information in tensor completion. In *Proceedings of the 25th International Conference Companion on World Wide Web*. 65–66.

- [24] Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. 2019. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory* 66, 3 (2019), 1760–1784.
- [25] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. 2012. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2012), 208–220.
- [26] Ian B Malone, David Cash, Gerard R Ridgway, David G MacManus, Sebastien Ourselin, Nick C Fox, and Jonathan M Schott. 2013. MIRIAD—Public release of a multiple time point Alzheimer’s MR imaging dataset. *NeuroImage* 70 (2013), 33–36.
- [27] Andrea Montanari and Nike Sun. 2018. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics* 71, 11 (2018), 2381–2425.
- [28] Atsuhiko Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. 2012. Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery* 25, 2 (2012), 298–324.
- [29] Madhav Nimishakavi, Bamdev Mishra, Manish Gupta, and Partha Talukdar. 2018. Inductive framework for multi-aspect streaming tensor completion with side information. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 307–316.
- [30] Aaron Potechin and David Steurer. 2017. Exact tensor completion with sum-of-squares. *arXiv preprint arXiv:1702.06237* (2017).
- [31] Devavrat Shah and Christina Lee Yu. 2019. Iterative collaborative filtering for sparse noisy tensor estimation. In *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 41–45.
- [32] Dogyoon Song, Christina E Lee, Yihua Li, and Devavrat Shah. 2016. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems*. 2155–2163.
- [33] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. 2010. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789* (2010).
- [34] Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. 2011. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems*. 972–980.
- [35] Dong Xia and Ming Yuan. 2017. On Polynomial Time Methods for Exact Low-Rank Tensor Completion. *Foundations of Computational Mathematics* (2017), 1–49.
- [36] Dong Xia, Ming Yuan, and Cun-Hui Zhang. 2017. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934* (2017).
- [37] Ming Yuan and Cun-Hui Zhang. 2016. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* 16, 4 (2016), 1031–1068.
- [38] Ming Yuan and Cun-Hui Zhang. 2017. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory* 63, 10 (2017), 6753–6766.
- [39] Dave Zachariah, Martin Sundin, Magnus Jansson, and Saikat Chatterjee. 2012. Alternating Least-Squares for Low-Rank Matrix Reconstruction. *IEEE Signal Processing Letters* 19, 4 (2012), 231–234. <https://doi.org/10.1109/LSP.2012.2188026>
- [40] Anru Zhang. 2019. Cross: Efficient low-rank tensor completion. *Ann. Statist.* 47, 2 (04 2019), 936–964. <https://doi.org/10.1214/18-AOS1694>
- [41] Zemin Zhang and Shuchin Aeron. 2016. Exact tensor completion using t-SVD. *IEEE Transactions on Signal Processing* 65, 6 (2016), 1511–1526.
- [42] Zemin Zhang, Gregory Ely, Shuchin Aeron, Ning Hao, and Misha Kilmer. 2014. Novel Methods for Multilinear Data Completion and De-noising Based on Tensor-SVD. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 3842–3849. <https://doi.org/10.1109/CVPR.2014.485>
- [43] Tengfei Zhou, Hui Qian, Zebang Shen, Chao Zhang, and Congfu Xu. 2017. Tensor completion with side information: A riemannian manifold approach. In *IJCAI*. 3539–3545.

7 APPENDIX A: PROOFS

7.1 Proof of Main Theorem

The proof of the main theorem uses Lemmas 2 and 1, whose proofs we defer to Sections 7.2 and 7.3. Given the results in these lemmas, the remaining proof of the main theorem follows from a straightforward analysis of the nearest neighbor averaging estimator given the estimated distances. *Proof of Theorem 1:* Recall that our final estimate is computed via a nearest neighbor,

$$\hat{T}(\mathbf{i}) = \frac{\sum_{\mathbf{i}'} T_3^{obs}(\mathbf{i}') K(\mathbf{i}, \mathbf{i}')}{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')} \text{ where } K(\mathbf{i}, \mathbf{i}') = \prod_{\ell \in [L]} \mathbb{1}_{[\hat{d}_\ell(i_\ell, i'_\ell) \leq \eta]}.$$

We can decompose the entrywise error into a bias term and an observation noise term,

$$|\hat{T}(\mathbf{i}) - T(\mathbf{i})| = \frac{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}') |T(\mathbf{i}') - T(\mathbf{i})|}{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')} + \left| \frac{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}') (T_3^{\text{obs}}(\mathbf{i}') - T(\mathbf{i}'))}{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')} \right|.$$

Let us denote $\rho_y(h) = \sum_{\mathbf{k} \in [r]^t} \Lambda(\mathbf{k}) \mathbb{1}_{[k_y=h]}$. The bias term we can bound by

$$\begin{aligned} |T(\mathbf{i}') - T(\mathbf{i})| &= \left| \sum_{\mathbf{k} \in [r]^t} \Lambda(\mathbf{k}) \left(\prod_{\ell \in [t]} q_{\ell k_\ell}(x_\ell(i_\ell)) - \prod_{\ell \in [t]} q_{\ell k_\ell}(x_\ell(i'_\ell)) \right) \right| \\ &\leq \sum_{\ell \in [t]} \left| \sum_{\mathbf{k} \in [r]^t} \Lambda(\mathbf{k}) \prod_{\ell' < \ell} q_{\ell' k_{\ell'}}(x_{\ell'}(i_{\ell'})) (q_{\ell k_\ell}(x_\ell(i_\ell)) - q_{\ell k_\ell}(x_\ell(i'_\ell))) \prod_{\ell'' > \ell} q_{\ell'' k_{\ell''}}(x_{\ell''}(i_{\ell''})) \right| \\ &\leq \sum_{\ell \in [t]} B^{t-1} \left| \sum_{h \in [r]} (q_{\ell h}(x_\ell(i_\ell)) - q_{\ell h}(x_\ell(i'_\ell))) \rho_\ell(h) \right| \\ &\leq \sum_{\ell \in [t]} B^{t-1} \sqrt{r \sum_{h \in [r]} (q_{\ell h}(x_\ell(i_\ell)) - q_{\ell h}(x_\ell(i'_\ell)))^2 \rho_\ell(h)^2} \\ &\leq B^{t-1} \sqrt{r} \sum_{\ell \in [t]} \max_{h \in [r]} |\rho_\ell(h)| \sqrt{\|Q_\ell^T(e_{i_\ell} - e_{i'_\ell})\|_2^2}, \end{aligned}$$

where we used the assumption that $\max_{\ell, k} \sup_x |q_{\ell k}(x)| \leq B$. By construction, $\hat{\Lambda} = \hat{U} \hat{\Sigma} \hat{V}^T$ and $\tilde{Q}_y = Q_y \hat{U}$ where $\hat{U}$ and $\hat{V}$ are orthonormal such that

$$\begin{aligned} \|\hat{\Sigma}^{s+1} \tilde{Q}_y^T(e_a - e_{a'})\|_2^2 &= (e_a - e_{a'})^T \tilde{Q}_y \hat{\Sigma}^{2(s+1)} \tilde{Q}_y^T(e_a - e_{a'}) \\ &= (e_a - e_{a'})^T Q_y (\hat{\Lambda} \hat{\Lambda}^T)^{(s+1)} Q_y(e_a - e_{a'}) \\ &\geq \hat{\sigma}_{\min}^{2(s+1)} \|Q_y^T(e_a - e_{a'})\|_2^2. \end{aligned}$$

Let $\rho_{\max} := \max_{\ell \in [t]} \max_{h \in [r]} |\rho_\ell(h)|$. It follows then that,

$$|T(\mathbf{i}') - T(\mathbf{i})| \leq B^{t-1} \sqrt{r} \rho_{\max} \hat{\sigma}_{\min}^{-(s+1)} \sum_{\ell \in [t]} \|\hat{\Sigma}^{s+1} \tilde{Q}_\ell^T(e_{i_\ell} - e_{i'_\ell})\|_2. \quad (28)$$

By construction, if $K(\mathbf{i}, \mathbf{i}') > 0$, then for all $\ell \in [t]$, $\hat{d}_\ell(i_\ell, i'_\ell) \leq \eta$. Conditioned on the good event from Lemma 1, it follows that for $\mathbf{i}, \mathbf{i}'$ such that $K(\mathbf{i}, \mathbf{i}') > 0$ and for all $\ell \in [t]$,

$$\|\hat{\Sigma}^{s+1} \tilde{Q}_\ell^T(e_{i_\ell} - e_{i'_\ell})\|_2^2 \leq \eta + \frac{5rB\hat{\sigma}_{\max}^{s+1}(1 + \hat{\sigma}_{\max}^s) \log^{1/2}(trn^2)(1 + o(1))}{(n\tilde{p})^{1/2}}.$$

Let $C_d = 5rB\hat{\sigma}_{\max}^{s+1}(1 + \hat{\sigma}_{\max}^s)$. Next we argue that the number of datapoints $\mathbf{i}' \in \Omega_3$ such that $K(\mathbf{i}, \mathbf{i}') > 0$ is sufficiently large such that the noise term in the estimate is small. Conditioned on the good event in Lemmas 1 and 2 with

$$\eta' = \eta - \frac{C_d \log^{1/2}(trn^2)(1 + o(1))}{(n\tilde{p})^{1/2}},$$

it follows that

$$\begin{aligned}
\sum_{\mathbf{i}'} K(\mathbf{i}, \mathbf{i}') &\geq \sum_{\mathbf{i}'} \prod_{\ell \in [t]} \mathbb{1} \left[\|\hat{\Sigma}^{s+1} \tilde{Q}_\ell^T (e_{i_\ell} - e_{i'_\ell})\|_2^2 \leq \eta - \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}} \right] \\
&= \prod_{\ell \in [t]} \left(1 + \sum_{\mathbf{i}' \neq \mathbf{i}_\ell} \mathbb{1} \left[\|\hat{\Sigma}^{s+1} \tilde{Q}_\ell^T (e_{i_\ell} - e_{i'_\ell})\|_2^2 \leq \eta - \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}} \right] \right) \\
&\geq \left(1 + \frac{(1-\delta)(n-1)}{\hat{\sigma}_{\max}^s L} \sqrt{\eta - \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}}} \right)^t. \tag{29}
\end{aligned}$$

We assumed that Ω_3 is a freshly sampled dataset such that $\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')$ is distributed as a Binomial($\sum_{\mathbf{i}'} K(\mathbf{i}, \mathbf{i}'), p$). By Chernoff's bound,

$$\mathbb{P} \left(\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}') \leq (1-\gamma)p \sum_{\mathbf{i}'} K(\mathbf{i}, \mathbf{i}') \right) \leq \exp \left(-\frac{1}{2} \gamma^2 p \sum_{\mathbf{i}'} K(\mathbf{i}, \mathbf{i}') \right). \tag{30}$$

Conditioned on Ω_3 , the noise terms $(T_3^{obs}(\mathbf{i}') - T(\mathbf{i}'))$ are independent, mean zero, and bounded in $[-1, 1]$. By Hoeffding's bound,

$$\mathbb{P} \left(\left| \frac{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}') (T_3^{obs}(\mathbf{i}') - T(\mathbf{i}'))}{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')} \right| \geq \sqrt{\frac{2 \log(n^{t+1})}{\sum_{\mathbf{i}'} \mathbb{1}_{[\mathbf{i}' \in \Omega_3]} K(\mathbf{i}, \mathbf{i}')}} \right) \leq \frac{2}{n^{t+1}}. \tag{31}$$

As a result, conditioned on the good events in Lemmas 1 and 2, with probability

$$1 - \frac{2}{n^{t+1}} - \exp \left(-\frac{1}{2} \gamma^2 p \left(1 + \frac{(1-\delta)(n-1)}{\hat{\sigma}_{\max}^s L} \sqrt{\eta - \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}}} \right)^t \right),$$

for constants δ, γ , and for

$$\eta \geq \frac{2C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}},$$

it holds that

$$\begin{aligned}
&|\hat{T}(\mathbf{i}) - T(\mathbf{i})| \\
&= tB^{t-1} \sqrt{r} \rho_{\max} \hat{\sigma}_{\min}^{-(s+1)} \sqrt{\eta + \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}}} \\
&\quad + \sqrt{2 \log(n^{t+1})} \left((1-\gamma)p \left(1 + \frac{(1-\delta)(n-1)}{\hat{\sigma}_{\max}^s L} \sqrt{\eta - \frac{C_d \log^{1/2}(trn^2)(1+o(1))}{(n\tilde{p})^{1/2}}} \right)^t \right)^{-\frac{1}{2}} \\
&= \Theta \left(tB^{t-1} \sqrt{r} \rho_{\max} \hat{\sigma}_{\min}^{-(s+1)} \sqrt{\max \left(\eta, \frac{rB \hat{\sigma}_{\max}^{2s+1} \log^{1/2}(n^2)}{n^{\kappa/2}} \right)} \right) + \Theta \left(\sqrt{\frac{\hat{\sigma}_{\max}^{st} L^t \log(n^{t+1})}{pn^t \eta^{t/2}}} \right).
\end{aligned}$$

By choosing η to balance the two terms, the bound is minimized for

$$\eta = \Theta \left(\max \left(\frac{\log^{1/2}(n^2)}{(n\tilde{p})^{1/2}}, \frac{\log^{2/(t+2)}(n^{t+1})}{n^{2(\kappa+1)/(t+2)}} \right) \right). \tag{32}$$

Recall that $n\tilde{p} = \Theta(n^{\min(\kappa, 1)})$. For this choice of η , it follows by (29) that with high probability

$$p \sum_{i'} K(\mathbf{i}, \mathbf{i}') = \Theta \left(p(n\sqrt{\eta'})^t \right) = \Omega \left(\frac{\log(n^{t+1})}{\eta} \right) = \Omega(n^{\min(\kappa/2, 1/2, 2(\kappa+1)/(t+2))}).$$

We can plug in this bound to simplify (30). Given this choice of η , we can also simplify the probability of error in Lemma 2 by the fact that

$$n\sqrt{\eta} = \Theta \left(\max \left(n^{1-\min(\kappa, 1)/4} \log^{1/4}(n^2), n^{1-(\kappa+1)/(t+2)} \log^{1/(t+2)}(n^{t+1}) \right) \right) = \Omega(n^{1/2}).$$

As we would like to show the entrywise error bound over all n^t entries, we take the intersection of all the good events and apply union bound. We use Lemma 1 for each of the $\ell \in [t]$ modes. We use 2 for each of the $\ell \in [t]$ modes and $i \in [n]$ coordinates. We union bound over all $\mathbf{i} \in [n]^t$ entries for the bounds in (30) and (31). It follows that for constant δ, γ , and η chosen according to (32), with probability at least

$$\begin{aligned} & 1 - \frac{6t(1+o(1))}{tn} - tn \exp \left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2\hat{\sigma}_{\max}^s L} \right) - n^t \exp \left(-\frac{1}{2}\gamma^2 p \sum_{i'} K(\mathbf{i}, \mathbf{i}') \right) - n^t \left(\frac{2}{n^{t+1}} \right) \\ &= 1 - \frac{8(1+o(1))}{n} - tn \exp \left(-\Omega(n^{1/2}) \right) - n^t \exp \left(-n^{\min(\kappa/2, 1/2, 2(\kappa+1)/(t+2))} \right) \\ &= 1 - \frac{8(1+o(1))}{n}, \end{aligned}$$

the estimate output by our algorithm satisfies

$$\max_{\mathbf{i} \in [n]^t} |\hat{T}(\mathbf{i}) - T(\mathbf{i})| = O \left(\max \left(\frac{\log^{1/4}(n^2)}{n^{\min(\kappa, 1)/4}}, \frac{\log^{1/(t+2)}(n^{t+1})}{n^{(\kappa+1)/(t+2)}} \right) \right).$$

It follows then that

$$\mathbb{E} \left[\frac{1}{n^t} \sum_{\mathbf{i} \in [n]^t} (\hat{T}(\mathbf{i}) - T(\mathbf{i}))^2 \right] = O \left(\max \left(\frac{\log^{1/2}(n^2)}{n^{\min(\kappa, 1)/2}}, \frac{\log^{2/(t+2)}(n^{t+1})}{n^{2(\kappa+1)/(t+2)}} \right) \right).$$

□

7.2 Sufficiently Many Nearest Neighbors

In this section, we prove Lemma 2, which states that there are sufficiently many nearest neighbors.

Proof of Lemma 2: By construction,

$$\begin{aligned} \|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2 &\leq \hat{\sigma}_{\max}^{2s} \|\hat{\Sigma} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2 \\ &= \hat{\sigma}_{\max}^{2s} \int_{0^1} (\tilde{f}(x_y(a), x') - \tilde{f}(x_y(a'), x'))^2 dx' \\ &\leq \hat{\sigma}_{\max}^{2s} L^2 (x_y(a) - x_y(a'))^2 \end{aligned}$$

where the last inequality follows from Lipschitzness of $\tilde{f}$ as shown in (27). As a result, for any $\eta' > 0$, if $|x_\ell(i) - x_\ell(i')| \leq \frac{\sqrt{\eta'}}{\hat{\sigma}_{\max}^s L}$, then $\|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2 \leq \eta'$. By the model assumption that for $i \in [n]$, the latent variables $x_\ell(i)$ are sampled i.i.d., it follows that for any $i \in [n]$,

$$\sum_{i' \neq i \in [n]} \mathbb{1}[\|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2 \leq \eta']$$

stochastically dominates a Binomial $\left(n-1, \frac{\sqrt{\eta'}}{\hat{\sigma}_{\max}^s L}\right)$ distributed random variable. Therefore the lemma statement follows by Chernoff's bound. □

7.3 Concentration of Distance Estimates

This section focuses on proving the key Lemma 1, which shows that the estimated distances concentrate well. We will assume $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$; it follows then that $s = \lceil \frac{\ln(n)}{\ln(pn^{t-1})} \rceil = \lceil \frac{1}{\kappa} \rceil$. When $\kappa < 1$, then $\tilde{p} = \Theta(n^{\kappa-1})$ such that $\tilde{p} = o(1)$. When $\kappa \geq 1$, then $\tilde{p} = \Theta(1)$ and the constructed matrix M_{yz}^{obs} is dense.

Let us denote event

$$\mathcal{A}_{a,s}^1(\delta, C) = \cup_{\ell=1}^{s-1} \{|\mathcal{S}_{a,\ell}| \in [((1-\delta)n\tilde{p})^\ell(1-o(1)), ((1+\delta)n\tilde{p})^\ell]\} \cup \{|\mathcal{S}_{a,s}| \geq Cn\} \cup \{|\mathcal{S}_{a,s+1}| \geq Cn\}.$$

LEMMA 3. Assume $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$. For $s = \lceil \frac{1}{\kappa} \rceil$ and any $\delta \in (0, 1)$ and $a \in [n]$, there exists some constant C such that

$$\mathbb{P}(\neg \mathcal{A}_{a,s}^1(C)) \leq 2 \exp\left(-\frac{\delta^2(n\tilde{p})(1-o(1))}{3}\right).$$

The proof is deferred to Section 7.4.

Recall that $\hat{\sigma}_k$ denotes the k -th singular value of the matrix $\hat{\Lambda}$ defined in (16), and $\hat{\Sigma}$ is the diagonal matrix with $\hat{\sigma}_k$ on the k -th diagonal entry.

Let us denote event

$$\mathcal{A}_{a,j,k}^2(\delta) = \begin{cases} \left\{ |e_k^T \tilde{Q}_y \tilde{N}_{a,j} - e_k^T \hat{\Sigma}^j \tilde{Q}_y e_a| \leq \frac{\hat{\sigma}_k^{j-1} \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}} \right\} & \text{if } j \text{ is even,} \\ \left\{ |e_k^T \tilde{Q}_z \tilde{N}_{a,j} - e_k^T \hat{\Sigma}^j \tilde{Q}_z e_a| \leq \frac{\hat{\sigma}_k^{j-1} \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}} \right\} & \text{if } j \text{ is odd.} \end{cases}$$

LEMMA 4. Assume $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$. For $s = \lceil \frac{1}{\kappa} \rceil$, constants $\delta \in (0, 1)$ and C , and any $a \in [n]$, $k \in [r]$, $j \in [s+1]$,

$$\mathbb{P}\left(\neg \mathcal{A}_{a,j,k}^2(\delta) \mid \mathcal{A}_{a,s}^1(\delta, C), \{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}\right) \leq \frac{2(1+o(1))}{trn^2},$$

where $\mathcal{T} = [t] \setminus \{y, z\}$.

The proof is deferred to Section 7.5.

If s is even, let us denote event

$$\begin{aligned} \mathcal{A}_{a,a'}^3 &= \left\{ \left| D(a, a') - \tilde{N}_{a,s}^T \tilde{Q}_y^T \hat{\Sigma} \tilde{Q}_z \tilde{N}_{a',s+1} \right| \right. \\ &\leq \max\left(\frac{2 \log^{1/2}(tn^3)}{(pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|)^{1/2}}, \frac{8 \log(tn^3)}{3|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|} \right) (1+o(1)) \left. \right\}. \end{aligned}$$

and if s is odd, let us denote event

$$\begin{aligned} \mathcal{A}_{a,a'}^3 &= \left\{ \left| D(a, a') - \tilde{N}_{a,s}^T \tilde{Q}_z^T \hat{\Sigma} \tilde{Q}_y \tilde{N}_{a',s+1} \right| \right. \\ &\leq \max\left(\frac{2 \log^{1/2}(tn^3)}{(pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|)^{1/2}}, \frac{8 \log(tn^3)}{3|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|} \right) (1+o(1)) \left. \right\}. \end{aligned}$$

LEMMA 5. Assume $p = n^{-(t-1)+\kappa}$ for $\kappa > 0$. For $s = \lceil \frac{1}{\kappa} \rceil$, constants $\delta \in (0, 1)$ and C , and any $a, a' \in [n]$, it holds that

$$\mathbb{P}\left(\neg \mathcal{A}_{a,a'}^3(\phi) \mid \mathcal{A}_{a,s}^1(\delta, C), \cap_{k=1}^r \mathcal{A}_{a,s,k}^2(\delta), \mathcal{A}_{a',s}^1(\delta, C), \cap_{k=1}^r \mathcal{A}_{a',s+1,k}^2(\delta), \{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}\right) \leq \frac{2}{tn^3} + \frac{1}{n^6}.$$

The proof is deferred to Section 7.6.

Proof of Lemma 1: Conditioned on events $\mathcal{A}_{a,s}^1(\delta, C), \mathcal{A}_{a',s}^1(\delta, C), \cap_{k=1}^r \mathcal{A}_{a,s,k}^2(\delta), \cap_{k=1}^r \mathcal{A}_{a',s+1,k}^2(\delta)$, if s is even,

$$\begin{aligned} & |\tilde{N}_{a,s}^T \tilde{Q}_y^T \hat{\Sigma} \tilde{Q}_z \tilde{N}_{a',s+1} - e_a^T \tilde{Q}_y^T \hat{\Sigma}^{2s+2} \tilde{Q}_y e_{a'}| \\ &= \left| \sum_k \hat{\sigma}_k (e_k^T \tilde{Q}_y \tilde{N}_{a,s}) (e_k^T \tilde{Q}_z \tilde{N}_{a',s+1} - e_k^T \hat{\Sigma}^{s+1} \tilde{Q}_y e_{a'}) + \sum_k \hat{\sigma}_k (e_k^T \tilde{Q}_y \tilde{N}_{a,s} - e_k^T \hat{\Sigma}^s \tilde{Q}_y e_a) (e_k^T \hat{\Sigma}^{s+1} \tilde{Q}_y e_{a'}) \right| \\ &\leq B \sum_k \frac{\hat{\sigma}_k^{s+1} \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}} + B \sum_k \frac{\hat{\sigma}_k^{2s+1} \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}} \\ &\leq \frac{rB\hat{\sigma}_{\max}^{s+1}(1+\hat{\sigma}_{\max}^s) \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}}. \end{aligned}$$

When s is odd, a similar argument shows that

$$|\tilde{N}_{a,s}^T \tilde{Q}_z^T \hat{\Sigma} \tilde{Q}_y \tilde{N}_{a',s+1} - e_a^T \tilde{Q}_y^T \hat{\Sigma}^{2s+2} \tilde{Q}_y e_{a'}| \leq \frac{rB\hat{\sigma}_{\max}^{s+1}(1+\hat{\sigma}_{\max}^s) \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}}.$$

To put it all together, conditioned on events $\mathcal{A}_{a,s}^1(\delta, C), \mathcal{A}_{a',s}^1(\delta, C), \cap_{k=1}^r (\mathcal{A}_{a,s,k}^2(\delta) \cap \mathcal{A}_{a',s+1,k}^2(\delta)), \cap_{k=1}^r (\mathcal{A}_{a',s,k}^2(\delta) \cap \mathcal{A}_{a',s+1,k}^2(\delta)), \mathcal{A}_{a,a'}^3(\phi), \mathcal{A}_{a,a}^3(\phi), \mathcal{A}_{a',a}^3(\phi), \mathcal{A}_{a',a'}^3(\phi)$, for constant δ ,

$$\begin{aligned} |\hat{d}_y(a, a') - \|\hat{\Sigma}^{s+1} \tilde{Q}_y^T (e_a - e_{a'})\|_2^2| &\leq \max\left(\frac{8 \log^{1/2}(tn^3)}{(pn^{t-2} |\mathcal{S}_{a,s}| |\mathcal{S}_{a',s+1}|)^{1/2}}, \frac{32 \log(tn^3)}{3 |\mathcal{S}_{a,s}| |\mathcal{S}_{a',s+1}|}\right) (1 + o(1)) \\ &\quad + \frac{4rB\hat{\sigma}_{\max}^{s+1}(1+\hat{\sigma}_{\max}^s) \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}} \end{aligned}$$

The first term scales as $\Theta\left(\max\left(\frac{\log^{1/2}(tn^3)}{n^{(1+\kappa)/2}}, \frac{\log(tn^3)}{n^2}\right)\right)$ and the second term scales as $\Theta\left(\frac{\log^{1/2}(trn^2)}{\min(n^{\kappa/2}, n^{1/2})}\right)$. The first term is always dominated by the second term. We plug in a choice of $\delta = \frac{9}{25}$ to get the final bound.

To guarantee this bound on the distance estimates for all pairs $a, a' \in [n]^2$, we take the intersection of the good events over pairs $a, a' \in [n]^2$. The max bound on the distance estimates hold when events $\cap_{a \in [n]} \mathcal{A}_{a,s}^1(\delta, C), \cap_{a \in [n]} \cap_{k=1}^r (\mathcal{A}_{a,s,k}^2(\delta) \cap \mathcal{A}_{a,s+1,k}^2(\delta)), \cap_{a,a' \in [n]^2} \mathcal{A}_{a,a'}^3(L)$ hold. By Lemmas 3, 4, 5 and union bound, for constant δ , these good events hold with probability at least

$$1 - 2n \exp\left(-\frac{\delta^2 n^\kappa (1 - o(1))}{3}\right) - \frac{4rn(1 + o(1))}{tn^2} - \frac{2n^2}{tn^3} - \frac{1}{n^6} = 1 - \frac{6(1 + o(1))}{tn}.$$

□

7.4 Rate of Local Neighborhood Growth

Proof of Lemma 3: We first handle the setting that $p = n^{-(t-1)+\kappa}$ for $\kappa \geq 1$. By definition,

$$\tilde{p} = (1 - (1 - p)^{n^{t-2}}) = (1 - (1 - \frac{n^{\kappa-1}}{n^{t-2}})^{n^{t-2}}).$$

Recall that $e^{-x} = \lim_{n \rightarrow \infty} (1 + \frac{-x}{n})^n$. Therefore for $\kappa = 1$, $\tilde{p} \rightarrow 1 - e^{-1} = \Theta(1)$. For $\kappa > 1$, then $\tilde{p} \rightarrow 1 = \Theta(1)$. As a result, in the setting where $\kappa \geq 1$, the density of observations $\tilde{p}$ in our constructed matrix is constant. For $s = 1$, $|\mathcal{S}_{a,s}| \sim \text{Binomial}(n, \tilde{p})$. By Chernoff's bound, it holds that for $C = (1 - \delta)\tilde{p} = \Theta(1)$,

$$\mathbb{P}(|\mathcal{S}_{a,s}| \geq Cn) \geq 1 - \exp\left(-\frac{\delta^2(n\tilde{p})}{3}\right).$$

Conditioned on $|\mathcal{S}_{a,s}| \geq Cn$, $|\mathcal{S}_{a,s+1}| \sim \text{Binomial}(n-1, 1-(1-\tilde{p})^{|\mathcal{S}_{a,s}|})$. As $1-\tilde{p} < 1$ and $|\mathcal{S}_{a,s}| = \omega(1)$, then $1 - (1-\tilde{p})^{|\mathcal{S}_{a,s}|} \rightarrow 1$. By Chernoff's bound, it holds that for $C = (1 - \delta)\tilde{p} = \Theta(1)$,

$$\mathbb{P}(|\mathcal{S}_{a,s+1}| \geq Cn \mid |\mathcal{S}_{a,s}| \geq Cn) \geq 1 - \exp\left(-\frac{(1-C)^2n}{3}\right).$$

For $\delta \in (0, 1)$, we can verify that $(1-C)^2 \geq \delta$.

We next address the ultra-sparse setting where $p = n^{-(t-1)+\kappa}$ for $\kappa \in (0, 1)$, such that $\tilde{p} = 1 - (1-p)^{n^{t-2}} = \Theta(n^{\kappa-1}) = o(1)$. Recall that our graph is bipartite between vertex sets $\mathcal{V}_1 = [n]$ and $\mathcal{V}_2 = [n]$. Without loss of generality, assume that $a \in \mathcal{V}_1$. Let $\mathcal{F}_{a,h}$ denote the sigma-algebra containing information about the latent parameters, edges and the values associated with vertices in the bipartite graph up to distance h from a , i.e. the depth h radius neighborhood of a .

Let $p_{a,\ell} = 1 - (1-\tilde{p})^{|\mathcal{S}_{a,\ell-1}|}$, and let

$$n_{a,\ell} = \begin{cases} |\mathcal{V}_2 \setminus \cup_{i=0}^{\lfloor \ell/2 \rfloor - 1} \mathcal{S}_{a,2i+1}| & \text{if } \ell \text{ is odd} \\ |\mathcal{V}_1 \setminus \cup_{i=0}^{\ell/2 - 1} \mathcal{S}_{a,2i}| & \text{if } \ell \text{ is even.} \end{cases}$$

For depth ℓ , conditioned on $\mathcal{F}_{\ell-1}$, $|\mathcal{S}_{a,\ell}|$ is distributed according to a Binomial with parameters $(n_{a,\ell}, p_{a,\ell})$. It follows by Chernoff's bound that

$$\mathbb{P}(|\mathcal{S}_{a,\ell}| \notin (1 \pm \delta)n_{a,\ell}p_{a,\ell} \mid \mathcal{F}_{a,\ell-1}) \leq 2 \exp\left(-\frac{\delta^2 n_{a,\ell} p_{a,\ell}}{3}\right).$$

Let us define the following event $\tilde{\mathcal{A}}_{a,h}$,

$$\tilde{\mathcal{A}}_{a,h} = \cap_{\ell=1}^h \{|\mathcal{S}_{a,\ell}| \notin (1 \pm \delta)n_{a,\ell}p_{a,\ell}\}.$$

Next we argue that for $s = \lceil \frac{1}{\kappa} \rceil$, event $\tilde{\mathcal{A}}_{a,s}$ implies that for all $\ell \in [s-1]$,

$$\{|\mathcal{S}_{a,\ell}| \in [((1-\delta)n\tilde{p})^\ell(1-o(1)), ((1+\delta)n\tilde{p})^\ell]\},$$

$|\mathcal{S}_{a,s}| = \Theta(n)$, and $|\mathcal{S}_{a,s+1}| = \Theta(n)$.

We first prove the upper bounds on $|\mathcal{S}_{a,\ell}|$. Naively, $|\mathcal{S}_{a,\ell}| \leq n$. Conditioned on $\tilde{\mathcal{A}}_{a,s}$,

$$\begin{aligned} |\mathcal{S}_{a,\ell}| &\leq (1+\delta)n_{a,\ell}p_{a,\ell} \\ &\leq (1+\delta)n \left(1 - (1-\tilde{p})^{|\mathcal{S}_{a,\ell-1}|}\right) \\ &\leq (1+\delta)n\tilde{p}|\mathcal{S}_{a,\ell-1}|. \end{aligned}$$

By inductively repeating this argument and using the fact that $|\mathcal{S}_{a,0}| = 1$, it follow that

$$|\mathcal{S}_{a,\ell}| \leq ((1+\delta)n\tilde{p})^\ell.$$

For $\ell < \frac{1}{\kappa}$, $(n\tilde{p})^\ell = o(n)$.

Next we prove the lower bounds on $|\mathcal{S}_{a,\ell}|$ for $\ell \in [s+1]$. Let us assume without loss of generality that ℓ is even (same argument holds for ℓ odd). Conditioned on $\tilde{\mathcal{A}}_{a,s}$,

$$\begin{aligned} n_{a,\ell} &\geq n - \sum_{i=0}^{\ell/2-1} |\mathcal{S}_{a,2i}| \\ &\geq n - \sum_{i=0}^{\ell/2-1} ((1+\delta)n\tilde{p})^{2i} \\ &= n(1 - o(1)) \end{aligned}$$

The last step follows from the fact that $2i \leq \ell - 2 \leq \lceil \frac{1}{\kappa} \rceil - 1 < \frac{1}{\kappa}$, such that $((1+\delta)n\tilde{p})^{2i} = o(1)$. Furthermore our choice of s guarantees it to be a constant (as κ is constant). Next we want to lower bound $p_{a,\ell}$. Conditioned on $\tilde{\mathcal{A}}_{a,s}$,

$$p_{a,\ell} = 1 - (1 - \tilde{p})^{|\mathcal{S}_{a,\ell-1}|} \geq \tilde{p}|\mathcal{S}_{a,\ell-1}|(1 - \tilde{p}|\mathcal{S}_{a,\ell-1}|)$$

This lower bound is only useful when $\tilde{p}|\mathcal{S}_{a,\ell-1}| = o(1)$, otherwise the bound could be negative. By the upper bound on $|\mathcal{S}_{a,\ell-1}|$, for $\ell < s$, it holds that $\tilde{p}|\mathcal{S}_{a,\ell-1}| \leq \frac{1}{n}(\tilde{p}n)^\ell = o(1)$. Therefore, for $\ell < s$,

$$\begin{aligned} |\mathcal{S}_{a,\ell}| &\geq (1 - \delta)n_{a,\ell}p_{a,\ell} \\ &\geq (1 - \delta)n(1 - o(1))\tilde{p}|\mathcal{S}_{a,\ell-1}|(1 - \tilde{p}|\mathcal{S}_{a,\ell-1}|) \\ &\geq (1 - \delta)n(1 - o(1))\tilde{p}|\mathcal{S}_{a,\ell-1}|(1 - o(1)). \end{aligned}$$

By inductively repeating this argument and using the fact that $|\mathcal{S}_{a,0}| = 1$, it follow that for $\ell < s$,

$$|\mathcal{S}_{a,\ell}| \geq ((1 + \delta)n\tilde{p})^\ell (1 - o(1)).$$

Next we prove that $|\mathcal{S}_{a,\ell}| = \Omega(n)$ for $\ell \in \{s, s+1\}$. Recall that $e^{-x} = \lim_{n \rightarrow \infty} (1 + \frac{-x}{n})^n$. For $\ell \geq s \geq \frac{1}{\kappa}$, using the fact that $\tilde{p} = \Theta(n^{\kappa-1})$,

$$\tilde{p}|\mathcal{S}_{a,\ell-1}| \geq \tilde{p}((1 + \delta)n\tilde{p})^{\frac{1-\kappa}{\kappa}} = \Theta(n^{\kappa-1}((1 + \delta)n^\kappa)^{\frac{1-\kappa}{\kappa}}) = \Theta(1), \text{ for some constant } > 0.$$

As a result,

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\tilde{p}|\mathcal{S}_{a,\ell-1}|}{|\mathcal{S}_{a,\ell-1}|}\right)^{|\mathcal{S}_{a,\ell-1}|} \leq e^{-\tilde{p}|\mathcal{S}_{a,\ell-1}|} < 1,$$

which implies that for $\ell \in \{s, s+1\}$, $p_{a,\ell} \geq C$ for some constant $C > 0$. Therefore, for $\ell < s$,

$$\begin{aligned} |\mathcal{S}_{a,\ell}| &\geq (1 - \delta)n_{a,\ell}p_{a,\ell} \\ &\geq (1 - \delta)n(1 - o(1))C \\ &= \Omega(n). \end{aligned}$$

To complete the proof, we use the lower bounds on $|\mathcal{S}_{a,\ell}|$ to reduce the probability bounds.

$$\begin{aligned} \mathbb{P}(\neg \tilde{\mathcal{A}}_{a,s+1}) &= \mathbb{P}(\cup_{\ell=1}^{s+1} \{|\mathcal{S}_{a,\ell}| \notin (1 \pm \delta)n_{a,\ell}p_{a,\ell}\}) \\ &\leq \sum_{\ell=1}^{s+1} \mathbb{P}\left(|\mathcal{S}_{a,\ell}| \notin (1 \pm \delta)n_{a,\ell}p_{a,\ell} \mid \tilde{\mathcal{A}}_{a,\ell-1}\right) \\ &\leq \sum_{\ell=1}^{s+1} 2 \exp\left(-\frac{\delta^2 n_{a,\ell} p_{a,\ell}}{3}\right) \\ &\leq \sum_{\ell=1}^{s-1} 2 \exp\left(-\frac{\delta^2 (n\tilde{p})^\ell (1 - o(1))}{3}\right) + 4 \exp\left(-\frac{\delta^2 nC}{3}\right). \end{aligned}$$

Note that the first term significantly dominates the remaining terms, as $n\tilde{p}$ is asymptotically smaller than n and $(n\tilde{p})^\ell$ for any $\ell > 1$. Therefore, the remaining terms get absorbed by the $o(1)$ in the exponent, leading to

$$\mathbb{P}(\neg \tilde{\mathcal{A}}_{a,s+1}) \leq 2 \exp\left(-\frac{\delta^2(n\tilde{p})(1-o(1))}{3}\right).$$

□

7.5 Martingale Concentration of Neighborhood Statistics

Proof of Lemma 4: Assume without loss of generality that $a \in \mathcal{V}_1$. Let us define

$$Y_{a,h} = \begin{cases} e_k^T \hat{\Sigma}^{s+1-h} \tilde{Q}_y^T \tilde{N}_{a,h} & \text{if } h \text{ is even} \\ e_k^T \hat{\Sigma}^{s+1-h} \tilde{Q}_z^T \tilde{N}_{a,h} & \text{if } h \text{ is odd} \end{cases}$$

$$D_{a,h} = Y_{a,h} - Y_{a,h-1}$$

so that $\sum_{h=1}^j D_{a,h} = Y_{a,j} - e_k^T \hat{\Sigma}^{s+1} \tilde{Q}_y^T e_a$. Conditioned on $\mathcal{S}_{a,\ell}$ for all $\ell \in [s+1]$ and conditioned on all $\{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}$, let $\mathcal{F}_{a,h}$ denote the sigma-algebra containing information about the latent parameters, edges and the values associated with vertices in the bipartite graph up to distance h from a , i.e. the depth h radius neighborhood of a . This includes $x_y(i)$ and $x_z(i)$ for all $i \in \cup_{\ell=0}^h \mathcal{S}_{a,h}$, as well as $M_{yz}^{obs}(i, j)$ for any edge i, j such that i or j is at distance at most $h-1$ from vertex a . Conditioned on $\mathcal{F}_{a,h}$, the BFS tree rooted at vertex a up to depth h is measurable, as are quantities $\tilde{N}_{a,\ell}$ for any $\ell \leq h$.

We will show that conditioned on $\mathcal{S}_{a,\ell}$ for all $\ell \in [s+1]$ and conditioned on all $\{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}$, $\{(D_{a,h}, \mathcal{F}_h)\}$ is a martingale difference sequence with controlled conditional variances such that martingale concentration holds.

Without loss of generality, let's assume that h is even (the below argument will also follow for odd h)

$$\begin{aligned} D_{a,h} &= e_k^T \hat{\Sigma}^{s+1-h} \tilde{Q}_y^T \tilde{N}_{a,h} - e_k^T \hat{\Sigma}^{s-h} \tilde{Q}_z^T \tilde{N}_{a,h-1} \\ &= \hat{\sigma}_k^{s+1-h} (e_k^T \tilde{Q}_y^T \tilde{N}_{a,h} - e_k^T \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a,h-1}) \\ &= \frac{\hat{\sigma}_k^{s+1-h}}{|\mathcal{S}_{a,h}|} \sum_{i \in \mathcal{S}_{a,h}} (N_{a,h}(i) \tilde{q}_{yk}(x_y(i)) - e_k^T \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a,h-1}). \end{aligned}$$

$D_{a,h}$ can be written as a sum of independent terms X_i for $i \in \mathcal{S}_{a,h}$,

$$\begin{aligned} X_i &= \frac{\hat{\sigma}_k^{s+1-h}}{|\mathcal{S}_{a,h}|} \left(N_{a,h}(i) \tilde{q}_{yk}(x_y(i)) - e_k^T \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a,h-1} \right) \\ &= \frac{\hat{\sigma}_k^{s+1-h}}{|\mathcal{S}_{a,h}|} \left(\sum_{b \in \mathcal{S}_{a,h-1}} N_{a,h-1}(b) \mathbb{1}_{[b=\pi(i)]} M_{yz}^{obs}(i, b) \tilde{q}_{yk}(x_y(i)) - e_k^T \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a,h-1} \right), \end{aligned}$$

where $\pi(i)$ denote the parent of i in the BFS tree. Conditioned on $\mathcal{F}_{u,h-1}$, for any $i \in \mathcal{S}_{a,h}$, any coordinate $b \in \mathcal{S}_{a,h-1}$ is equally likely to be the parent of i in the BFS tree due to the symmetry/uniformity of the sampling process. As a result, conditioned on $\mathcal{F}_{u,h-1}$ and $i \in \mathcal{S}_{a,h}$,

$$\mathbb{E}[X_i] = \frac{\hat{\sigma}_k^{s+1-h}}{|\mathcal{S}_{a,h}| |\mathcal{S}_{a,h-1}|} \sum_{b \in \mathcal{S}_{a,h-1}} N_{a,h-1}(b) \left(\mathbb{E} \left[M_{yz}^{obs}(i, b) \tilde{q}_{yk}(x_y(i)) \mid (i, b) \in \tilde{\Omega}_{yz} \right] - \tilde{q}_{zk}(x_z(b)) \hat{\sigma}_k \right).$$

First we verify that

$$\begin{aligned}
 & \mathbb{E} \left[\frac{\mathbb{1}_{[\mathbf{h} \in \Omega_1]}}{|\Omega_1 \cap \mathcal{I}_{yz}(i, b)|} \mid (i, b) \in \tilde{\Omega}_{yz} \right] \\
 &= \sum_{g=0}^{n^{t-2}-1} \frac{1}{1+g} \mathbb{P}(\mathbf{h} \in \Omega_1, |\Omega_1 \cap \mathcal{I}_{yz}(i, b)| = g+1 \mid |\Omega_1 \cap \mathcal{I}_{yz}(i, b)| \geq 1) \\
 &= \sum_{y=0}^{n^{t-2}-1} \frac{1}{1+y} \frac{p^{\binom{n^{t-2}-1}{g}} p^g (1-p)^{n^{t-2}-1-g}}{1 - (1-p)^{n^{t-2}}} \\
 &= \frac{p}{1 - (1-p)^{n^{t-2}}} \sum_{g=0}^{n^{t-2}-1} \frac{1}{p n^{t-2} \binom{n^{t-2}}{g+1}} p^{g+1} (1-p)^{n^{t-2}-(g+1)} \\
 &= \frac{p}{1 - (1-p)^{n^{t-2}}} \frac{1 - (1-p)^{n^{t-2}}}{p n^{t-2}} \\
 &= \frac{1}{n^{t-2}}.
 \end{aligned}$$

Conditioned on $\{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y, z\}}$, i.e. all latent variables in modes $3, 4, \dots, t$, by (23),

$$\begin{aligned}
 \mathbb{E} \left[M_{yz}^{obs}(i, b) \tilde{q}_{yk}(x_y(i)) \mid (i, b) \in \tilde{\Omega}_{yz} \right] &= \sum_{h \in [r]} \hat{\sigma}_h \tilde{q}_{yh}(x_y(i)) \tilde{q}_{yk}(x_y(i)) \tilde{q}_{zh}(x_z(b)) \\
 &= \hat{\sigma}_k \tilde{q}_{zk}(x_z(b)),
 \end{aligned}$$

implying that $\mathbb{E}[X_i] = 0$. Furthermore, as $\|N_{a,h}\|_\infty \leq 1$,

$$|X_i| \leq \frac{\tilde{B} \hat{\sigma}_k^{s+1-h} (1 + |\hat{\sigma}_k|)}{|\mathcal{S}_{a,h}|}.$$

Therefore $\{(D_{a,h}, \mathcal{F}_h)\}$ is a martingale difference sequence with uniformly bounded differences. Next we want to establish concentration. Using the model assumptions that $|T^{obs}(\mathbf{i})| \leq 1$ such that $|M_{yz}^{obs}(i, b)| \leq 1$ and $\|N_{a,h}\|_\infty \leq 1$, it follows that

$$\begin{aligned}
 \text{Var}[X_i \mid i \in \mathcal{S}_{a,h}] &= \frac{\hat{\sigma}_k^{2(s+1-h)}}{|\mathcal{S}_{a,h}|^2} \text{Var} \left[\sum_{b \in \mathcal{S}_{a,h-1}} N_{a,h-1}(b) \mathbb{1}_{[b=\pi(i)]} M_{yz}^{obs}(i, b) \tilde{q}_{yk}(x_y(i)) \right] \\
 &\leq \frac{\hat{\sigma}_k^{2(s+1-h)}}{|\mathcal{S}_{a,h}|^2} \mathbb{E} \left[\sum_{b \in \mathcal{S}_{a,h-1}} \mathbb{1}_{[b=\pi(i)]} N_{a,h-1}^2(b) (M_{yz}^{obs}(i, b))^2 (\tilde{q}_{yk}(x_y(i)))^2 \right] \\
 &\leq \frac{\hat{\sigma}_k^{2(s+1-h)}}{|\mathcal{S}_{a,h}|^2} \mathbb{E} \left[\sum_{b \in \mathcal{S}_{a,h-1}} \frac{1}{|\mathcal{S}_{a,h-1}|} \right] \\
 &= \frac{\hat{\sigma}_k^{2(s+1-h)}}{|\mathcal{S}_{a,h}|^2}.
 \end{aligned}$$

Conditioned on $\{S_{a,\ell}\}_{\ell \in [s+1]}$ and $\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}$, it follows that $D_{u,h}$ conditioned on $\mathcal{F}_{u,h-1}$ is sub-exponential with parameters

$$\left(\frac{\hat{\sigma}_k^{(s+1-h)}}{\sqrt{|S_{a,h}|}}, \frac{\tilde{B}\hat{\sigma}_k^{s+1-h}(1+|\hat{\sigma}_k|)}{|S_{a,h}|} \right).$$

Conditioned on the event $\mathcal{A}_{a,s}^1(\delta, C)$, the quantity $|S_{a,h}|$ can be lower bounded so that the sub-exponential parameters are bounded above by

$$\left(\frac{\hat{\sigma}_k^{(s+1-h)}}{\sqrt{((1-\delta)n\tilde{p})^h(1-o(1))}}, \frac{\tilde{B}\hat{\sigma}_k^{s+1-h}(1+|\hat{\sigma}_k|)}{((1-\delta)n\tilde{p})^h(1-o(1))} \right)$$

for $h \in [s-1]$, and

$$\left(\frac{\hat{\sigma}_k^{(s+1-h)}}{\sqrt{Cn}}, \frac{\tilde{B}\hat{\sigma}_k^{s+1-h}(1+|\hat{\sigma}_k|)}{Cn} \right)$$

for $h \in \{s, s+1\}$.

By the Bernstein style bound for martingale concentration, it holds that $\sum_{h=1}^j D_{a,j}$ is sub-exponential with parameters ν_* and α_* for

$$\begin{aligned} \nu_* &= \sqrt{\sum_{h=1}^{\min(s-1,j)} \frac{\hat{\sigma}_k^{2(s+1-h)}}{((1-\delta)n\tilde{p})^h(1-o(1))} + \frac{\hat{\sigma}_k^2 \mathbb{1}_{[s \leq j]}}{Cn} + \frac{\mathbb{1}_{[s+1 \leq j]}}{Cn}} \\ &= \frac{\hat{\sigma}_k^s(1+o(1))}{((1-\delta)n\tilde{p})^{1/2}} \\ \alpha_* &= \max \left(\max_{h \in [s-1]} \frac{\tilde{B}\hat{\sigma}_k^{s+1-h}(1+|\hat{\sigma}_k|)}{((1-\delta)n\tilde{p})^h(1-o(1))}, \frac{\tilde{B}\hat{\sigma}_k(1+|\hat{\sigma}_k|)}{Cn}, \frac{\tilde{B}(1+|\hat{\sigma}_k|)}{Cn} \right) \\ &= \frac{\tilde{B}\hat{\sigma}_k^s(1+|\hat{\sigma}_k|)(1+o(1))}{(1-\delta)n\tilde{p}} \end{aligned}$$

where we use the fact that for sufficiently large n , $((1-\delta)n\tilde{p})^{-1}$ asymptotically dominates $((1-\delta)n\tilde{p})^{-h}$ for any $h > 1$. For the setting where $\kappa \geq 1$ and $s = 1$, we choose constant $C = (1-\delta)\tilde{p}$ such that ν_* and α_* also scale as the expressions above. It follows by Bernstein's inequality that for $0 < z < \frac{\hat{\sigma}_k^s(1-o(1))}{\tilde{B}(1+|\hat{\sigma}_k|)}$,

$$\mathbb{P} \left(|Y_{a,j} - e_k^T \hat{\Sigma}^{s+1} \tilde{Q}_y^T e_a| \geq \theta \mid \mathcal{A}_{a,s}^1(\delta, C), \{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}} \right) \leq 2 \exp \left(- \frac{(1-\delta)n\tilde{p}\theta^2(1-o(1))}{2\hat{\sigma}_k^{2s}} \right).$$

We will choose $\theta = \hat{\sigma}_k^s((1-\delta)n\tilde{p})^{-1/2} \log^{1/2}(trn^2)$, such that with probability $1 - \frac{2(1+o(1))}{trn^2}$,

$$|Y_{a,j} - e_k^T \hat{\Sigma}^{s+1} \tilde{Q}_y^T e_a| \leq \frac{\hat{\sigma}_k^s \log^{1/2}(trn^2)}{((1-\delta)n\tilde{p})^{1/2}},$$

which implies event $\mathcal{A}_{a,j}^2(\delta)$ holds. $\square$

7.6 Concentration of Inner Product Statistic

Proof of Lemma 5: Recall that we assume T_2^{obs} is a fresh data sample (alternatively this assumption can be removed by sample splitting instead). Let us define the shrink $_\phi$ operator to be

$$\text{shrink}_\phi(x) = \begin{cases} \phi & \text{if } x > \phi \\ x & \text{if } x \in [-\phi, \phi] \\ -\phi & \text{if } x < -\phi \end{cases}.$$

Without loss of generality assume s is even (the same argument follows for odd s). Let us define

$$X_{ij} = N_{a,s}(i)N_{a',s+1}(j) \left(\sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} T_2^{obs}(\mathbf{h}) \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(h_\ell) \right)$$

$$\tilde{X}_{ij} = N_{a,s}(i)N_{a',s+1}(j) \text{shrink}_\phi \left(\sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} T_2^{obs}(i,j,h) \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(h_\ell) \right).$$

The statistic $D(a, a')$ as defined in (9) can be constructed as sums of X_{ij} . We will show that with high probability $X_{ij} = \tilde{X}_{ij}$, and additionally $|\mathbb{E}[X_{ij}] - \mathbb{E}[\tilde{X}_{ij}]|$ is small. As $|T_2^{obs}(\mathbf{h})| \leq 1$ and $|W_\ell(h_\ell)| \leq 1$ by our model assumptions, $|\sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} T_2^{obs}(\mathbf{h}) \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(h_\ell)| \leq |\Omega_2 \cap \mathcal{I}_{yz}(i,j)|$. It follows that $X_{ij} = \tilde{X}_{ij}$ whenever $|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| \leq \phi$. The difference between their expected values can be expressed as

$$|\mathbb{E}[X_{ij}] - \mathbb{E}[\tilde{X}_{ij}]| \leq \mathbb{1}_{[i \in S_{a,s}, j \in S_{a',s+1}]} \mathbb{E} \left[\mathbb{1}_{[|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| > \phi]} (|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| - \phi) \right],$$

where $|\Omega_2 \cap \mathcal{I}_{yz}(i,j)|$ is distributed as Bernoulli(n^{t-2}, p).

For the setting where $\kappa \geq 1$, by Chernoff's bound,

$$\mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| > \phi) \leq \exp \left(-\frac{\phi - pn^{t-2}}{3} \right)$$

and

$$\begin{aligned} \mathbb{E} \left[\mathbb{1}_{[|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| > \phi]} (|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| - \phi) \right] &= \sum_{g=1}^{\infty} \mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| \geq \phi + g) \\ &\leq \sum_{g=1}^{\infty} \exp \left(-\frac{\phi + g - pn^{t-2}}{3} \right) \\ &\leq 3 \exp \left(-\frac{\phi - pn^{t-2}}{3} \right). \end{aligned}$$

If $\kappa = 1$, we choose $\phi = 24 \log(n)$ such that $\mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| > \phi)$ and $|\mathbb{E}[X_{ij}] - \mathbb{E}[\tilde{X}_{ij}]|$ are bounded by $O(n^{-8})$. If $\kappa > 1$, then we choose $\phi = 2pn^{t-2}$ such that the $\mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i,j)| > \phi)$ and $|\mathbb{E}[X_{ij}] - \mathbb{E}[\tilde{X}_{ij}]|$ decays exponentially in n , i.e. bounded by $\exp(-n^{\kappa-1})$.

For the ultra-sparse setting where $\kappa \in (0, 1)$, we use a different argument as Chernoff's bound is not strong enough.

$$\begin{aligned}
\mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i, j)| > \phi) &= \sum_{g=\phi+1}^{n^{t-2}} \binom{n^{t-2}}{g} p^g (1-p)^{n^{t-2}-g} \\
&< (1-p)^{n^{t-2}} \sum_{g=\phi+1}^{n^{t-2}} \left(\frac{pn^{t-2}}{1-p} \right)^g \\
&< (1-p)^{n^{t-2}} \left(\frac{pn^{t-2}}{1-p} \right)^{\phi+1} \left(1 - \frac{pn^{t-2}}{1-p} \right)^{-1} \\
&< \frac{(1-p)^{n^{t-2}+1}}{1-p(n^{t-2}+1)} \left(\frac{pn^{t-2}}{1-p} \right)^{\phi+1} \\
&= (pn^{t-2})^{\phi+1} (1+o(1)) = n^{-(1-\kappa)(\phi+1)} (1+o(1)).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E} \left[\mathbb{1}_{[|\Omega_2 \cap \mathcal{I}_{yz}(i, j)| > \phi]} (|\Omega_2 \cap \mathcal{I}_{yz}(i, j)| - \phi) \right] &= \sum_{g=\phi+1}^{n^{t-2}} (g - \phi) \binom{n^{t-2}}{g} p^g (1-p)^{n^{t-2}-g} \\
&\leq (1-p)^{n^{t-2}} \sum_{g=1}^{n^{t-2}-\phi} g \left(\frac{pn^{t-2}}{1-p} \right)^{g+\phi} \\
&\leq (1-p)^{n^{t-2}} \left(\frac{pn^{t-2}}{1-p} \right)^{\phi+1} \left(1 - \frac{pn^{t-2}}{1-p} \right)^{-2} \\
&= \frac{(1-p)^{n^{t-2}+2}}{(1-p(n^{t-2}+1))^2} \left(\frac{pn^{t-2}}{1-p} \right)^{\phi+1}.
\end{aligned}$$

We choose $\phi = \lceil 8 \ln(n) / \ln((pn^{t-2})^{-1}) \rceil = \lceil 8/(1-\kappa) \rceil$, so that $\mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i, j)| > \phi)$ and $|\mathbb{E}[X_{ij}] - \mathbb{E}[\tilde{X}_{ij}]|$ are bounded by $O(n^{-8})$. Therefore, by union bound, $\mathbb{P}(\cup_{ij} \{X_{ij} \neq \tilde{X}_{ij}\}) \leq \sum_{ij} \mathbb{P}(|\Omega_2 \cap \mathcal{I}_{yz}(i, j)| \geq \phi) \leq n^{-6}$.

Next, we show that $\tilde{X}_{ij}$ concentrates around $\mathbb{E}[\tilde{X}_{ij}]$. As $\tilde{X}_{ij}$ results from shrinking values of X_{ij} towards zero, the variance of $\tilde{X}_{ij}$ is bounded by the variance of X_{ij} . Conditioned on $\mathcal{F}_{a,s}, \mathcal{F}_{a',s+1}$ and $\{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}$, i.e. the latent variables for modes $t \setminus \{y, z\}$,

$$\begin{aligned}
&\text{Var}[X_{ij} \mid \mathcal{F}_{a,s}, \mathcal{F}_{a',s+1}, \{\{x_\ell(i)\}_{i \in [n]}\}_{\ell \in [t] \setminus \{y,z\}}] \\
&= N_{a,s}^2(i) N_{a',s+1}^2(j) \sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} \sum_{\mathbf{h}' \in \mathcal{I}_{yz}(i,j)} \text{Cov}[T_2^{\text{obs}}(\mathbf{h}), T_2^{\text{obs}}(\mathbf{h}')] \\
&= N_{a,s}^2(i) N_{a',s+1}^2(j) \sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} \text{Var}[T_2^{\text{obs}}(\mathbf{h})] \\
&\leq pn^{t-2} N_{a,s}^2(i) N_{a',s+1}^2(j).
\end{aligned}$$

By independence of $\tilde{X}_{ij}$ and given that $\|N_{a,s}\|_\infty \leq 1$ and $\|N_{a',s+1}\|_\infty \leq 1$,

$$\text{Var} \left[\sum_{i,j} \tilde{X}_{ij} \right] \leq pn^{t-2} |\mathcal{S}_{a,s}| |\mathcal{S}_{a',s+1}|.$$

By construction of the shrink operator, $|\tilde{X}_{ij}| \leq \phi$. As a result, it follows by Bernstein's inequality that

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i,j} (\tilde{X}_{ij} - \mathbb{E} [\tilde{X}_{ij}]) \right| \geq \theta \mid \mathcal{F}_{a,s}, \mathcal{F}_{a',s+1}, \{ \{x_\ell(i)\}_{i \in [n]} \}_{\ell \in [t] \setminus \{y,z\}} \right) \\ & \leq 2 \exp \left(- \frac{3\theta^2}{6pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}| + 2\phi\theta} \right) \end{aligned}$$

We choose $\theta = \max \left(2 \log^{1/2}(tn^3)(pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|)^{1/2}, \frac{4\phi \log(tn^3)}{3} \right)$, such that the above probability is bounded above by $2/tn^3$. Conditioned on $\mathcal{A}_{a,s}^1(\delta, C)$, $\mathcal{A}_{a',s}^1(\delta, C)$, it follows that $|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}| = \Theta(n)$ such that $pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}| = \Theta(n^{1+\kappa})$.

Conditioned on $\mathcal{F}_{a,s}$, $\mathcal{F}_{a',s+1}$ and $\{ \{x_\ell(i)\}_{i \in [n]} \}_{\ell \in [t] \setminus \{y,z\}}$,

$$\begin{aligned} & \mathbb{E} \left[X_{ij} \mid \mathcal{F}_{a,s}, \mathcal{F}_{a',s+1}, \{ \{x_\ell(i)\}_{i \in [n]} \}_{\ell \in [t] \setminus \{y,z\}} \right] \\ & = N_{a,s}(i)N_{a',s+1}(j) \left(\sum_{\mathbf{h} \in \mathcal{I}_{yz}(i,j)} \mathbb{E} [T_2^{obs}(\mathbf{h})] \prod_{\ell \in [t] \setminus \{y,z\}} W_\ell(h_\ell) \right) \\ & = N_{a,s}(i)N_{a',s+1}(j)p \sum_{\mathbf{k}} \Lambda(\mathbf{k})q_{yk_y}(x_y(i))q_{zk_z}(x_z(j)) \prod_{\ell \in [t] \setminus \{y,z\}} \sum_{h \in [n]} q_{\ell k_\ell}(x_\ell(h))W_\ell(h) \\ & = pN_{a,s}(i)N_{a',s+1}(j) \sum_{a,b \in [r]^2} q_{ya}(x_y(i))q_{zb}(x_z(j))\hat{\Lambda}(a,b)n^{t-2} \\ & = pn^{t-2}N_{a,s}(i)N_{a',s+1}(j)e_i^T Q_y \hat{\Lambda} Q_z^T e_j \\ & = pn^{t-2}N_{a,s}(i)N_{a',s+1}(j)e_i^T \tilde{Q}_y \hat{\Sigma} \tilde{Q}_z^T e_j \end{aligned}$$

such that

$$\frac{1}{pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|} \sum_{i,j} \mathbb{E} \left[X_{ij} \mid \mathcal{F}_{a,s}, \mathcal{F}_{a',s+1}, \{ \{x_\ell(i)\}_{i \in [n]} \}_{\ell \in [t] \setminus \{y,z\}} \right] = \tilde{N}_{a,s}^T \tilde{Q}_y \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a',s+1}.$$

For sufficiently large n , conditioned on $\mathcal{A}_{a,s}^1(\delta, C)$, $\mathcal{A}_{a',s}^1(\delta, C)$, $\mathcal{A}_{a,s}^2(\delta)$, $\mathcal{A}_{a',s+1}^2(\delta)$ for constants δ and C , it holds that with probability $1 - \frac{2}{tn^3} - \frac{1}{n^6}$,

$$\begin{aligned} & |D(a, a') - \tilde{N}_{a,s}^T \tilde{Q}_y \hat{\Sigma} \tilde{Q}_z^T \tilde{N}_{a',s+1}| \\ & \leq \max \left(\frac{2 \log^{1/2}(tn^3)}{(pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|)^{1/2}}, \frac{4\phi \log(tn^3)}{3pn^{t-2}|\mathcal{S}_{a,s}||\mathcal{S}_{a',s+1}|} \right) (1 + o(1)) \\ & = \Theta \left(\max \left(\frac{\log^{1/2}(tn^3)}{n^{(1+\kappa)/2}}, \frac{\log(tn^3)}{n^2} \right) \right), \end{aligned}$$

where the last equality comes from plugging in the choice of ϕ as

$$\phi = \begin{cases} \lceil 8 \ln(n)/\ln((pn^{t-2})^{-1}) \rceil & \text{if } pn^{t-2} = n^{\kappa-1} \text{ for } \kappa \in (0, 1) \\ 24 \log(n) & \text{if } pn^{t-2} = \Theta(1) \\ 2pn^{t-2} & \text{if } pn^{t-2} = n^{\kappa-1} \text{ for } \kappa > 1, \end{cases}$$

and verifying the above holds for each case of $\kappa < 1$, $\kappa = 1$, and $\kappa > 1$.

□

Received February 2022; revised March 2022; accepted April 2022