Robust Max Entrywise Error Bounds for Tensor Estimation From Sparse Observations via Similarity-Based Collaborative Filtering

Devavrat Shah, *Fellow, IEEE*, and Christina Lee Yu, *Member, IEEE*

***Abstract*— Consider the task of estimating a 3-order $n \times n \times n$ tensor from noisy observations of randomly chosen entries in the sparse regime. We introduce a similarity based collaborative filtering algorithm for estimating a tensor from sparse observations and argue that it achieves sample complexity that nearly matches the conjectured computationally efficient lower bound on the sample complexity for the setting of low-rank tensors. Our algorithm uses the matrix obtained from the flattened tensor to compute similarity, and estimates the tensor entries using a nearest neighbor estimator. We prove that the algorithm recovers a finite rank tensor with maximum entry-wise error (MEE) and mean-squared-error (MSE) decaying to 0 as long as each entry is observed independently with probability $p = \Omega(n^{-3/2+\kappa})$ for any arbitrarily small $\kappa > 0$. More generally, we establish robustness of the estimator, showing that when arbitrary noise bounded by $\varepsilon \geq 0$ is added to each observation, the estimation error with respect to MEE and MSE degrades by $\text{poly}(\varepsilon)$. Consequently, even if the tensor may not have finite rank but can be approximated within $\varepsilon \geq 0$ by a finite rank tensor, then the estimation error converges to $\text{poly}(\varepsilon)$. Our analysis sheds insight into the conjectured sample complexity lower bound, showing that it matches the connectivity threshold of the graph used by our algorithm for estimating similarity between coordinates.**

***Index Terms*— Sparse observations, tensor estimation, max entrywise error bounds, low rank models, latent variable models, collaborative filtering, nearest neighbor estimator.**

I. Introduction

TENSOR estimation involves the task of predicting underlying structure in a high-dimensional tensor structured dataset given only a sparse subset of observations. We call this "tensor estimation" rather than the conventional "tensor completion" as the goal is not only to fill missing entries but also to estimate entries whose noisy observations are available. Whereas matrices represent data associated to two modes, rows and columns, tensors represent data associated to general d modes. For example, a datapoint collected from a user-product interaction an e-commerce platform may be associated to a user, product, and date/time, which could be represented in a 3-order tensor where the three modes would correspond to users, products, and date/time. Image data is also naturally represented in a 3-order tensor format, with two modes representing the location of the pixel, and the third mode representing the RGB color components. Video data furthermore introduces a fourth mode indexing the time. Dynamic network data can also be represented in a tensor with one mode indexing the time and the other two modes indexing the nodes in the network.

There are many applications in which the dataset inherently has a lot of noise or is very sparsely observed. For example, e-commerce data is typically very sparse as the typical number of products a user interacts with is very small relative to the entire product catalog; furthermore the timepoints at which the user interacts with the platform may be sparse. When the dataset can be represented as a matrix, equivalent to a 2-order tensor, there has been a significant amount of research in designing practical algorithms and studying statistical limits for matrix estimation, a critical step in data pre-processing. Under conditions on uniform sampling and incoherence, the minimum sample complexity for estimation has been tightly characterized and achieved by simple algorithms. It is a natural and relevant question then to consider whether the techniques developed can extend to higher order tensors as well.

The previous literature has primarily focused on attaining consistency with respect to the mean squared error (MSE). Unfortunately as this is aggregated over the error in the full tensor, it does not translate to consistent bounds on entrywise error, as the error on a single entry could be very large despite the MSE being small due to averaging over many entries. However, entrywise bounds are important in practice as the results of tensor estimation are often used subsequently for decisions that involve comparisons between the estimates of individual entries.

In this work we focus on attaining consistent max entrywise error bounds by extending similarity based collaborative filtering algorithms to tensor estimation. Similarity based collaborative filtering is widely used in industry due to its simplicity, interpretability, and amenability to distributed and parallelized implementations. In the analysis of our proposed

Manuscript received 10 September 2021; revised 25 October 2022; accepted 10 January 2023. Date of publication 16 January 2023; date of current version 21 April 2023. This work was supported by NSF under Grant CCF-1948256 and Grant CNS-1955997. The work of Christina Lee Yu was supported by an Intel Rising Stars Award. An earlier version of this paper was presented in part at the 2019 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT.2019.8849683] and in part at the 2019 Allerton [DOI: 10.1109/ALLERTON.2019.8919933]. *(Corresponding author: Christina Lee Yu.)*

Devavrat Shah is with the Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Christina Lee Yu is with the Department of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: cleeyu@cornell.edu).

Communicated by M. Davenport, Associate Editor for Signal Processing and Information Theory.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3237231.

Digital Object Identifier 10.1109/TIT.2023.3237231

algorithm we show that it achieves a sample complexity that nearly matches a conjectured lower bound for computationally efficient algorithms. Perhaps most notably, our theoretical guarantees provide high probability bounds on the maximum entrywise error of the estimate, which is significantly stronger than the typical mean squared error style bounds found in the literature for other algorithms. We also provide error bounds under arbitrary bounded noise, which has implications towards approximately low rank settings.

A. Related Literature

Algorithms for analyzing sparse low rank matrices (equivalent to 2-order tensors) where the observations are sampled uniformly randomly have been well-studied. The algorithms consist of spectral decomposition or matrix factorization [1], [2], [3], nuclear norm minimization [4], [5], [6], [7], [8], [9], gradient descent [1], [2], [10], [11], [12], alternating minimization [13], [14], and nearest neighbor style collaborative filtering [15], [16], [17], [18], [19]. These algorithms have been shown to be provably consistent as long as the number of observations is $\Omega(rn\,\mathsf{poly}(\log n))$ for the noiseless setting where r is the rank and n is the number of rows and columns [1], [4]; similar results have been attained under additive Gaussian noise [2], [5] and generic bounded noise [3], [18]. Lower bounds show that $\Omega(rn)$ samples are necessary for consistent estimation, and $\Omega(rn\log(n))$ samples are necessary for exact recovery [5], [6], implying that the proposed algorithms are nearly sample efficient order-wise up to the information theoretic lower bounds.

There are results extending matrix estimation algorithms to higher order tensor estimation, assuming the tensor is low rank and that observations are sampled uniformly at random. The earliest approaches simply flatten or unfold the tensor to a matrix and subsequently apply matrix estimation algorithms [20], [21], [22], [23]. A d-order tensor where each dimension is length n would be unfolded to a $n^{\lfloor d/2\rfloor}\times n^{\lceil d/2\rceil}$ matrix, resulting in a sample complexity of $O(n^{\lceil d/2\rceil}\,\mathsf{poly}(\log n))$, significantly larger than the natural statistical lower bound that is linear with n due to the model being parameterized by linear in n latent variables. When d is odd, for example $d=3$ the sample complexity for this naive approach scales as $O(n^2\,\mathsf{poly}(\log n))$.

Subsequent works have improved upon this sample complexity, requiring only $\Omega(n^{3/2}\,\mathsf{poly}(\log n))$ observed entries for a 3-order tensor [24], [25], [26], [27], [28], [29], [30], [31]. References [24] and [25] analyzes the alternating minimization algorithm for exact recovery of the tensor given noiseless observations and finite rank $r=\Theta(1)$. References [27] and [28] use the sum of squares (SOS) method, and [29] introduces a spectral method. Both of these latter algorithms can handle noisy observations and overcomplete tensors where the rank is larger than the dimension. References [30] and [31] furthermore characterize the minimax optimal rate for the MSE and achieve it using spectral initialization followed by power iteration. For a general d-order tensor these results translate into a sample complexity scaling as $O(n^{d/2})$, improving upon $O(n^{\lceil d/2\rceil}$. Reference [32] prove that tensor nuclear norm minimization can recover the underlying low-rank d-order tensor with $O(n^{3/2}\,\mathsf{poly}(\log n))$ samples in the noiseless setting; however, the algorithm is not efficiently computable as computing tensor nuclear norm is NP-hard [33].

Reference [27] conjecture that any polynomial time estimator for a 3-order tensor must require $\Omega(n^{3/2})$ samples, based on a reduction between tensor estimation for a rank-1 tensor to the random 3-XOR distinguishability problem. They argue that if using the sum of squares hierarchy to construct relaxations for tensor rank, any result that achieves a consistent estimator with fewer than $n^{3/2}$ samples will violate a conjectured hardness of random 3-XOR distinguishability. Information theoretic bounds imply that one needs at least $\Omega(drn)$ observations to recover a d-order rank r tensor, consistent with the degrees of freedom or number of parameters in the model. Interestingly, this implies a conjectured gap between the computational and statistically achievable sample complexities, highlighting how tensor estimation is distinctly more difficult than matrix estimation.

The majority of the results in tensor estimation provide bounds on the mean squared error, which aggregates errors across entries. In contrast our results will also provide bounds on the maximum entrywise error. There has been recent interest on developing matrix estimation methods that provide max entrywise bounds using a leave one out analysis, cf. [34], [35], [36], [37], [38], [39]. Subseqeuntly [40], [41] extended the leave one out analysis to the tensor setting to obtain entrywise error bounds for a gradient descent algorithm with spectral initialization. The analysis and algorithm in our paper is significantly different than their work, as it results from showing high probability guarantees on the similarity computation, which is akin to a spectral algorithm, followed by a nearest neighbor analysis. Our results suggests that a combination of spectral analysis and nearest neighbor smoothing can achieve entrywise consistent estimates without further gradient descent refinements. As nearest neighbor methods are still widely used in industry, understanding their theoretical performance is of interest.

B. Contribution

Our results answer the following unresolved questions in the literature.

1. Is there a computationally efficient estimator that can provide a consistent estimation of low-rank tensor with respect to maximum entry-wise error (MEE) with minimal sample complexity of $\Omega(n^{\frac{3}{2}})$ in the presence of noise?
2. Is there an extension of matrix estimation collaborative filtering algorithm for the setting of tensor estimation that can provide consistent estimation with such minimal sample complexity?
3. Can the estimator be robust to adversarial bounded noise in the observations?

To begin with, we propose an algorithm for a symmetric 3-order tensor estimation which generalizes a nearest neighbor collaborative filtering algorithm for sparse matrix estimation introduced in [18]. Naïvely applying the matrix estimation

algorithm in [18] to the $n \times n^2$ matrix obtained by unfolding the 3-order tensor would require $\Omega(n^2)$ samples, far more than the desired sample complexity of $\Omega(n^{\frac{3}{2}})$. However, we argue that such a matrix obtained from the unfolded tensor can be used, after non-trivial modification, to compute the similarities between rows accurately using $\Omega(n^{\frac{3}{2}+\kappa})$ samples for any positive $\kappa > 0$. After computing these similarities we can achieve consistent estimation via a nearest neighbor estimator by additionally using the tensor structure.

Specifically, we establish that the mean squared error (MSE) in the estimation converges to 0 as long as $\Omega(n^{3/2+\kappa})$ random samples are observed for any $\kappa > 0$ for tensor with rank $r = \Theta(1)$. We further establish a stronger guarantee that the maximum entry-wise error (MEE) converge to 0 with high probability with similar sample complexity of $\Omega(n^{3/2+\kappa})$. Thus, this simple iterative collaborative filtering algorithm nearly achieves the conjectured computational sample complexity lower bound of $\Omega(n^{3/2})$ for tensor estimation. While we present the results for symmetric tensors, our method and analysis can extend to asymmetric tensors, which we discuss in Section V-D.

Beyond low-rank tensors, our results hold for tensors with potentially countably infinite rank as long as they can be well approximated by a low-rank tensor. Specifically, if the tensor can be approximated with $\varepsilon \geq 0$ with respect to max-norm by a rank $r = \Theta(1)$ tensor, then the MSE converges to $\mathsf{poly}(\varepsilon)$ and MEE converges to $\mathsf{poly}(\varepsilon)$ with high probability as long as $\Omega(n^{3/2+\kappa})$ random samples are observed for any $\kappa > 0$. This follows as a consequence of the robustness property of the algorithm that we establish: if arbitrary noise bounded by $\varepsilon \geq 0$ is added to each observation, then the estimation error with respect to MEE and MSE degrades by $\mathsf{poly}(\varepsilon)$.

To establish our results, the key analytic tool is utilizing certain concentration properties of a bilinear form arising from the local neighborhood expansion of any given coordinate for an asymmetric matrix with dimensions $n \times n^2$. This generalizes the analysis of a similar property for symmetric matrices in the prior work of [18]. Specifically, establishing the desired concentration requires handling dependencies arising in the local neighborhood expansion of the 3-order tensor that was absent in the matrix setting considered in [18]. Subsequently, we require a novel analytic method compared to the prior work. In particular we believe that the proof techniques in Lemma 7.7 may be useful to other settings in which one may desire a tighter concentration on sums of sparse random variables. As a consequence, we also establish performance guarantees for matrix estimation for asymmetric matrices having dimensions of different order, generalizing beyond of [18].

The algorithm and analysis also sheds insight on the conjectured lower bound for 3-order tensor. In particular, the threshold of $n^{3/2}$ is precisely the density of observations needed for the connectivity in the associated graph that is utilized to calculate similarities. If the graph is disconnected, the similarities can not be computed, while if the graph is connected, we are able to show that similarity calculations yield an excellent estimator. Understanding this relationship further remains an interesting open research direction.

A benefit of our algorithm is that it can be implemented in a parallelized manner where the similarities between pair of indices are computed in parallel. This lends itself to a distributed, scalable implementation. A naive bound on the computational complexity of our algorithm for 3-order tensor is at most pn^6. As discussed in Section V-D, with use of approximate nearest neighbors, these can be further improved and made truly implementable.[1]

II. Preliminaries

Tensor estimation from sparse observations hinges on an assumption that the true model exhibits low dimensional structure despite the high dimensional representation. However, there is not a unique definition of rank in the tensor setting, as natural generalizations of matrix rank lead to different quantities when extended to higher order tensors. We will focus on two commonly used definitions of tensor rank, the CP rank and the Tucker or multilinear rank.

For a d-order tensor $F \in \mathbb{R}^{n^d}$, we can decompose F into a sum of rank-1 tensors. For example if $d = 3$, then

$$F = \sum_{k=1}^{r} u_k \otimes v_k \otimes w_k,$$

where $\{u_k, v_k, w_k\}_{k\in[r]}$ is a collection of length n vectors. The CP-rank is the minimum number r such that F can be written as a sum of r rank-1 tensors, which we refer to as a CP-decomposition. The CP-rank may in fact be larger than the dimension n, and furthermore the latent vectors need not be orthogonal as is the case in the matrix setting.

An alternate notion of tensor rank is defined according to the dimension of subspaces corresponding to each mode. Let $F_{(y)}$ denote the unfolded tensor along the y-th mode, which is a matrix of dimension $n \times n^{d-1}$. Let columns of $F_{(y)}$ be referred to as mode y fibers of tensor F as depicted in Figure 1. The Tucker rank, or multilinear rank, is a vector $(r_1, r_2, \ldots r_d)$ such that for each mode $\ell \in [d]$, r_ℓ is the dimension of the column space of $F_{(y)}$. The Tucker rank is also the minimal values of $(r_1, r_2, \ldots r_d)$ such that the tensor F can be decomposed according to a multilinear multiplication of a core tensor $\Lambda \in \mathbb{R}^{r_1 \times r_2 \times \ldots r_d}$ with latent factor matrices $Q_1 \ldots Q_d$ for $Q_\ell \in \mathbb{R}^{n_\ell \times r_\ell}$, denoted as

$$\begin{aligned} F &= (Q_1 \otimes \cdots Q_d) \cdot (\Lambda) \\ &:= \sum_{\mathbf{k}\in[r_1]\times[r_2]\cdots\times[r_d]} \Lambda(\mathbf{k}) Q_1(\cdot, k_1) \otimes Q_2(\cdot, k_2) \cdots \otimes Q_d(\cdot, k_d), \end{aligned} \tag{II.1}$$

[1] A weaker abbreviated version of this result appeared in [42] without any proofs or discussion. Since the preliminary results, the convergence rates of the error have improved, and we have new results showing a perturbation analysis under arbitrary bounded noise, which extends our results to the approximately low rank setting. We also present a modification of the algorithm that significantly improves the overall computational complexity of the algorithm.

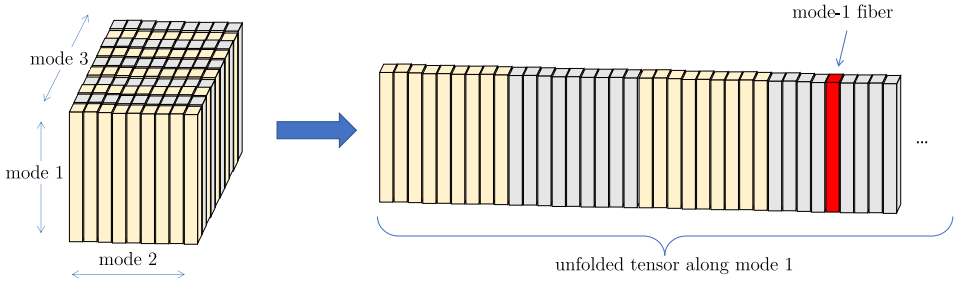

Fig. 1. Depicting an unfolding of a 3rd order tensor along mode 1. The columns of the resulting matrix are referred to as the mode-1 fibers of the tensor.

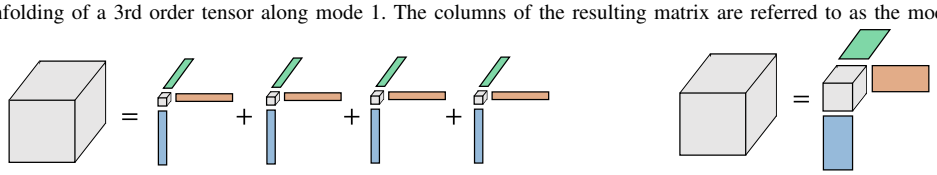

Fig. 2. (Left) The tensor CP-rank admits a decomposition corresponding to the sum of r rank-1 tensors. (Right) The Tucker rank or multilinear rank $(r_1, r_2, \dots r_d)$ admits a decomposition corresponding to a multilinear multiplication of a core tensor of dimensions $(r_1, r_2, \dots r_d)$ with latent factor matrices associated to each mode.

and depicted in Figure 2. The higher order SVD (HOSVD) specifies a unique Tucker decomposition in which the factor matrices $Q_1 \dots Q_d$ are orthonormal and correspond to the left singular vectors of the unfolded tensor along each mode [43].

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 $(r_1, r_2, \dots r_d)$, the CP-rank is bounded by the number of nonzero entries in the core tensor, which is at most $r_1 r_2 \cdots r_d / (\max_\ell r_\ell)$ [43]. 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 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 [27]. Our results also extend beyond to general tensors as well, though the presentation is simpler in the orthogonal setting.

III. Problem Statement and Model

Consider an $n \times n \times n$ symmetric tensor F generated as follows: For each $u \in [n]$, sample $\theta_u \sim U[0,1]$ independently. Let the true underlying tensor F be described by a Lipschitz function f evaluated over the latent variables, $F(u,v,w) = f(\theta_u, \theta_v, \theta_w)$ for $u, v, w \in [n]$. Without loss of generality, we shall assume that $\sup_{u,v,w \in [0,1]} |f(\theta_u, \theta_v, \theta_w)| \leq 1$. For example, if the coordinates of one mode of the tensor represent users or products in an e-commerce platform, one can view the latent variables associated to a coordinate as the unknown "type" of the user/product, which can be thought of as sampled i.i.d. from an underlying population distribution. The latent function f would then describe the expected observed interaction between units of type θ_u, θ_v, and θ_w.

Let M denote the observed symmetric data tensor, and let $\Omega \subseteq [n]^3$ denote the set of observed indices. Due to the symmetry, it is sufficient to restrict the index set to triplets (u,v,w) such that $u \leq v \leq w$, as the datapoint is identical for all other permutations of the same triplet. The datapoint at each of these distinct triplets $\{(u,v,w) : u \leq v \leq w\}$ is observed independently with probability $p \in (0,1]$, where we assume the observation is corrupted by mean zero independent additive noise terms. For $(u,v,w) \in \Omega$,

$$M(u,v,w) = F(u,v,w) + \epsilon_{uvw}, \tag{III.1}$$

and for $(u,v,w) \notin \Omega$, $M(u,v,w) = \star$.[2] We shall assume that $|M(u,v,w)| \leq 1$ with probability 1. We allow ϵ_{uvw} to have different distributions for different distinct triplets (u,v,w) as long as it is uniformly bounded so as to satisfy the boundedness constraints on $|M(u,v,w)|$ and $|F(u,v,w)|$. When the observations $M(u,v,w)$ are binary, this model is equivalent to the 3-uniform simple lipschitz hypergraphon in [44], which states a generative model for hypergraphs where the hyperedges consist of size 3 vertex sets. In this setting, $F(u,v,w) \in [0,1]$ would represent the probability of observing the hyperedge (u,v,w), and $M(u,v,w) \in \{0,1\}$ would indicate the presence of the hyperedge (u,v,w). The noise term ϵ_{uvw} is clearly bounded since the observations are binary. For any application in which the observations $M(u,v,w)$ are bounded, then $F(u,v,w) = \mathbb{E}[M(u,v,w)]$ would also be bounded, such that the boundedness on the noise ϵ_{uvw} would be reasonable. However, for applications in which $M(u,v,w)$ may not be bounded, or the almost sure bound is very large, we can extend our analysis to allow for sub-Gaussian noise ϵ_{uvw} rather than uniformly bounded noise, which is further discussed in section VI.

The goal is to recover the underlying tensor F from the incomplete noisy observation M so that the mean squared error (MSE) is small, where MSE for an estimate $\hat{F}$ is defined as

$$\text{MSE}(\hat{F}) := \mathbb{E}\left[\frac{1}{n^3} \sum_{(u,v,w) \in [n]^3} (\hat{F}(u,v,w) - F(u,v,w))^2\right]. \tag{III.2}$$

[2]The notation of $\star$ is used to denote the missing observation. When convenient, we shall replace $\star$ by 0 for the purpose of computation.

We will also be interested in the maximum entry-wise error (MEE) defined as

$$\|F - \hat{F}\|_{\max} := \max_{(u,v,w)\in[n]^3} |\hat{F}(u,v,w) - F(u,v,w)|. \tag{III.3}$$

A. Finite Spectrum

Consider the setting where the function f has finite spectrum. That is,

$$f(u,v,w) = \sum_{k=1}^{r} \lambda_k q_k(\theta_u) q_k(\theta_v) q_k(\theta_w),$$

where $r = \Theta(1)$ and $q_k(.)$ denotes the orthonormal ℓ_2 eigenfunctions, satisfying $\int_0^1 q_k(\theta)^2 d\theta = 1$ and $\int_0^1 q_k(\theta) q_h(\theta) d\theta = 0$ for $k \neq h$. Assume that the eigenfunctions are bounded, i.e. $|q_k(\theta)| \leq B$ for all $k \in [r]$.

Let Λ denote the diagonal $r \times r$ matrix where $\Lambda_{kk} = \lambda_k$. Let Q denote the $r \times n$ matrix where $Q_{ka} = q_k(\theta_a)$. Let $\mathcal{Q}$ denote the $r \times \binom{n}{2}$ matrix where $\mathcal{Q}_{kb} = q_k(\theta_{b_1}) q_k(\theta_{b_2})$ for some $b \in \binom{n}{2}$ that represents the pair of vertices (b_1, b_2) for $b_1 < b_2$. The finite spectrum assumption for f implies that the sampled tensor F is such that,

$$F = \sum_{k=1}^{r} \lambda_k (Q^T e_k) \otimes (Q^T e_k) \otimes (Q^T e_k).$$

That is, F has CP-rank at most r. In above and in the remainder of the paper, e_k denotes a vector with all 0s but kth entry being 1 of appropriate dimension (here it is r).

B. Approximately Finite Spectrum

In general, f may not have finite spectrum, e.g. a generic analytic function f. For such a setting, we shall consider f with approximately finite spectrum. Specifically, a function $f : [0,1]^3 \to \mathbb{R}$, it is said to have ε-approximate finite spectrum with rank r for $\varepsilon \geq 0$ if there exists a symmetric function $f_r : [0,1]^3 \to \mathbb{R}$ such that

$$\sup_{\theta_u,\theta_v,\theta_w\in[0,1]} |f(\theta_u,\theta_v,\theta_w) - f_r(\theta_u,\theta_v,\theta_w)| \leq \varepsilon$$
$$F_r(u,v,w) = f_r(\theta_u,\theta_v,\theta_w) = \sum_{k=1}^{r} \lambda_k q_k(\theta_u) q_k(\theta_v) q_k(\theta_w), \tag{III.4}$$

where $r = \Theta(1)$ and $q_k(.)$ denotes the orthonormal ℓ_2 eigenfunctions as before. That is, they satisfy $\int_0^1 q_k(\theta)^2 d\theta = 1$, $\int_0^1 q_k(\theta) q_h(\theta) d\theta = 0$ for $k \neq h$ and $|q_k(\theta)| \leq B$ for all $k \in [r]$.

The above describe property of f implies that the sampled tensor F is has ε-approximate rank r such that $F_r = \sum_{k=1}^{r} \lambda_k (Q^T e_k) \otimes (Q^T e_k) \otimes (Q^T e_k)$ and

$$\|F - F_r\|_{\max} \leq \varepsilon.$$

C. Extensions Beyond Orthogonal CP-Rank

The orthogonality conditions on our latent variable decomposition imply that the tensor F can be written as a sum of r rank-1 tensors, where the latent factors are approximately orthogonal. Alternately, this would suggest a Tucker decomposition of the tensor where the core tensor is superdiagonal. Not all tensors admit an orthogonal CP-rank decomposition, but this assumption has also been used in the literature as in [24]. This assumption of the existence of an orthogonal CP-rank decomposition can be relaxed as the main property that our algorithm and analysis use is the orthogonal decomposition of the unfolded tensors along each mode. Our algorithm and analysis will still extend to tensors with Tucker rank bounded by (r,r,r). For a general (r,r,r) Tucker rank tensor, we would instead carry out the analysis with respect to the latent orthogonal factors corresponding to the SVD of the unfolded tensor into a matrix, and the algorithm would use the same procedure to estimate similarities along each of the three modes separately. The presentation is stated for the orthogonal symmetric setting for simplicity.

D. Comparision of Assumptions With Literature

In the decomposition of the model f when it has finite spectrum, we assume that the functions q_k are orthonormal. This induces a decomposition of tensor F in terms of $Q \in \mathbb{R}^{r\times n}$ with respect to the sampled latent features $\theta \sim U[0,1]$. The rank r of the underlying decomposition is assumed to be $\Theta(1)$. We compare and contrast these with those assumed in the tensor estimation literature.

Most literature on tensor estimation do not impose a distribution on the underlying latent variables, but instead assume deterministic 'incoherence' style conditions on the latent singular vectors associated to the underlying tensor decomposition. This plays a similar role to our combined assumption of q_k being orthonormal and the latent variables sampled from a uniform distribution so that the mass in the singular vector matrix is roughly uniformly spread. For example, the notion of incoherence used in [27] imposes that the entries of the latent factors are bounded by a constant when the norm of the latent factor vectors scales as $\Theta(\sqrt{n})$. As $\theta_u \sim U[0,1]$, it holds that the latent factor vector $(q_k(\theta_1), q_k(\theta_2), \ldots q_k(\theta_n))$ will have norm scaling as $\sqrt{n}$. Due to the boundedness assumption that $|q_k(\theta)| \leq B$, our model will satisfy incoherence as defined in [27]. Some of the literature on tensor estimation allows for overcomplete tensors, i.e. $r > n$. While our finite spectrum setup requires $r = \Theta(1)$, the approximately finite spectrum can allow for potentially countably infinite spectrum but with sharply decaying spectrum so that it has ε-approximate rank being $r = \Theta(1)$.

In order to establish our result for the approximately finite spectrum setting, we perform a perturbation analysis wherein each observed entry is perturbed arbitrarily bounded by ε in magnitude: we shall establish that the resulting estimation error is changed by $\text{poly}(\varepsilon)$, both with respect to the MSE and max-norm. That is, with respect to arbitrary bounded noise in the observations, we are able to characterize the error induced by our method, which is of interest in its own right.

We remark on the Lipschitz property of f: the Lipschitz assumption implies that the tensor is "smooth", and thus there are sets of rows and columns that are similar to one another. As our algorithm is based on a nearest neighbor style approach we need that for any coordinate u, there is a significant mass of other coordinates a that are similar to u with respect to the function behavior. Other regularity conditions beyond

Lipschitz that would also guarantee sufficiently many "nearest neighbors" would lead to similar results for our algorithm. Lipschitzness also implies approximate low rankness as a Lipschitz function can be approximated by a piecewise constant function, where the number of pieces would then upper bound the rank.

IV. Algorithm

The algorithm is a nearest neighbor style in which the first phase is to estimate a distance function between coordinates, denoted $\mathsf{dist}(u,a)$ for all $(u,a) \in [n]^2$. Given the similarities, for some threshold η, the algorithm estimates by averaging datapoints from coordinates (a,b,c) for which $\mathsf{dist}(u,a) \le \eta$, $\mathsf{dist}(v,b) \le \eta$, and $\mathsf{dist}(w,c) \le \eta$.

The entry $F(a,b,c)$ depends on a coordinate a through its representation in the eigenspace, given by Qe_a. Therefore $f(a,b,c) \approx f(u,v,w)$ as long as $Qe_u \approx Qe_a$, $Qe_v \approx Qe_b$, and $Qe_w \approx Qe_c$. Ideally we would like our distance function $\mathsf{dist}(u,a)$ to approximate $\|Qe_u - Qe_a\|_2$, but these are hidden latent features that we do not have direct access to.

Let's start with a thought experiment supposing that the density of observations were $p = \omega(n^{-1})$ and the noise variance is σ^2 for all entries. For a pair of coordinates u and a, the expected number of pairs (b,c) such that both (u,b,c) and (a,b,c) are observed is on the order of $p^2 n^2 = \omega(1)$. For fixed θ_a, θ_u, and for randomly sampled θ_b, θ_c, the expected squared difference between the two corresponding datapoints reflects the distance between Qe_a and Qe_u along with the overall level of noise,

$$\begin{aligned}
&\mathbb{E}[(M(a,b,c) - M(u,b,c))^2 \mid \theta_a, \theta_u] \\
&= \mathbb{E}[(F(a,b,c) - F(u,b,c))^2 \mid \theta_a, \theta_u] + \mathbb{E}[\epsilon_{abc}^2 + \epsilon_{ubc}^2] \\
&= \mathbb{E}[(\textstyle\sum_k \lambda_k (q_k(\theta_a) - q_k(\theta_u)) q_k(\theta_b) q_k(\theta_c))^2 \mid \theta_a, \theta_u] + 2\sigma^2 \\
&= \mathbb{E}[\textstyle\sum_k \lambda_k^2 (q_k(\theta_a) - q_k(\theta_u))^2 q_k(\theta_b)^2 q_k(\theta_c)^2 \mid \theta_a, \theta_u] + 2\sigma^2 \\
&= \textstyle\sum_k \lambda_k^2 (q_k(\theta_a) - q_k(\theta_u))^2 + \sigma^2 \\
&= \|\Lambda Q(e_a - e_u)\|_2^2 + 2\sigma^2,
\end{aligned}$$

where we use the fact that $q_k(\cdot)$ are orthonormal. This suggests that approximating $\mathsf{dist}(u,a)$ with the average squared difference between datapoints corresponding to pairs (b,c) for which both (u,b,c) and (a,b,c) are observed.

This method does not attain the $p = n^{-3/2}$ sample complexity, as the expected number of pairs (b,c) for which (a,b,c) and (u,b,c) are both observed will go to zero for $p = o(n^{-1})$. This limitation arises due to the fact that when $p = o(n^{-1})$, the observations are extremely sparse. Consider the $n \times \binom{n}{2}$ "flattened" matrix of the tensor where row u correspond to coordinates $u \in [n]$, and columns correspond to pairs of indices, e.g. $(b,c) \in [n] \times [n]$ with $b \le c$. For any given row u, there are very few other rows that share observations along any column with the given row u, i.e. the number of 'neighbors' of any row index is few. If we wanted to exploit the intuition of the above simple calculations, we have to somehow enrich the neighborhood. We do so by constructing a graph using the non-zero pattern of the matrix as an adjacency matrix. This mirrors the idea from [18] for matrix estimation, which approximates distances by comparing expanded depth $2t+1$ local neighborhoods in the graph representing the sparsity pattern of the unfolded or flattened tensor. In particular, we will construct a statistic $\mathsf{dist}(u,a)$ such that with high probability it concentrates around $d(u,a)$ for

$$d(\theta_u, \theta_a) = \|\Lambda^{2t+1} Q(e_u - e_a)\|_2^2 = \sum_{k=1}^{r} \lambda_k^{4t+2} (q_k(\theta_u) - q_k(\theta_a))^2. \tag{IV.1}$$

As $F(u,v,w) = f(\theta_u, \theta_v, \theta_w) = \sum_{k=1}^r \lambda_k q_k(\theta_u) q_k(\theta_v) q_k(\theta_w)$, we can show that if $d(\theta_u, \theta_a)$, $d(\theta_v, \theta_b)$, and $d(\theta_w, \theta_c)$ are small, then $F(u,v,w)$ will be close to $F(a,b,c)$. The remaining challenge thus how to approximate $d(u,a)$. Consider a length $2t$ path in the bipartite graph from u to a, denoted by $(u, e^1, x^1, e^2, x^2, \ldots x^{t-1}, e^t, a)$, where $x^1, \ldots x^{t-1}$ are distinct coordinates in $[n] \setminus \{u,a\}$, and $e^1, \ldots e^t$ consist of pairs in $[n]^2$ such that the coordinates represented in these pairs are distinct from each other as well as $\{u, a, x^1, \ldots x^{t-1}\}$. Let us denote the pair $e^i = (e_1^i, e_2^i)$. Then the product of weights along this path in expectation conditioned on θ_u, θ_a is equal to $e_u^T Q^T \Lambda^{2t} Q e_a$ as shown in (IV.2) shown at the bottom of the next page.

Therefore, the product of weights along the path connecting u to a is a good proxy of quantity $e_u^T Q^T \Lambda^{2t} Q e_a$, for paths which do not revisit coordinates. The algorithm first constructs the local neighborhood of depth $2t$ centered at each coordinate u, and then connects these neighborhoods to form paths of length $4t+2$ in total, where t is chosen such that for every u, a there are sufficiently many paths used to construct the statistic $\mathsf{dist}(u,a)$ to guarantee concentration. The tensor setting requires an important modification of how one constructs the local breadth-first-search (BFS) trees due to the shared latent variables across different modes, as described in step 3 below.

A. Formal Description

We provide a formal description of the algorithm below. The crux of the algorithm is to compute similarity between any pair of indices using the matrix obtained by flattening the tensor, and then using a nearest neighbor estimator using these similarities between indices over the tensor structure. Details are as follows.

Step 1: Sample Splitting: Let us assume for simplicity of the analysis that we obtain 2 independent fresh observation sets of the data, Ω_1 and Ω_2. Tensors M_1 and M_2 contain information from the subset of the data in M associated to Ω_1 and Ω_2 respectively. M_1 is used to compute pairwise similarities between coordinates, and M_2 is used to average over datapoints for the final estimate. Furthermore, we take the coordinates $[n]$ and split it into two sets, $[n] = \{1, 2, \ldots, n/2\} \cup \{n/2+1, n/2+2, \ldots n\}$. Without loss of generality, let's assume that n is even. Let $\mathcal{V}_A$ denote the set of coordinate pairs within set 1 consisting of distinct coordinates, i.e. $\mathcal{V}_A = \{(b,c) \in [n/2]^2 \; s.t. \; b < c\}$. Let $\mathcal{V}_B$ denote the set of coordinate pairs within set 2 consisting of distinct coordinates, i.e. $\mathcal{V}_B = \{(b,c) \in ([n] \setminus [n/2])^2 \; s.t. \; b < c\}$. The sizes of $|\mathcal{V}_A|$ and $|\mathcal{V}_B|$ are both equal to $\binom{n/2}{2}$. We define M_A to be the n-by-$\binom{n/2}{2}$ matrix taking values $M_A(a,(b,c)) = M_1(a,b,c)$, where each row corresponds to an original coordinate of the

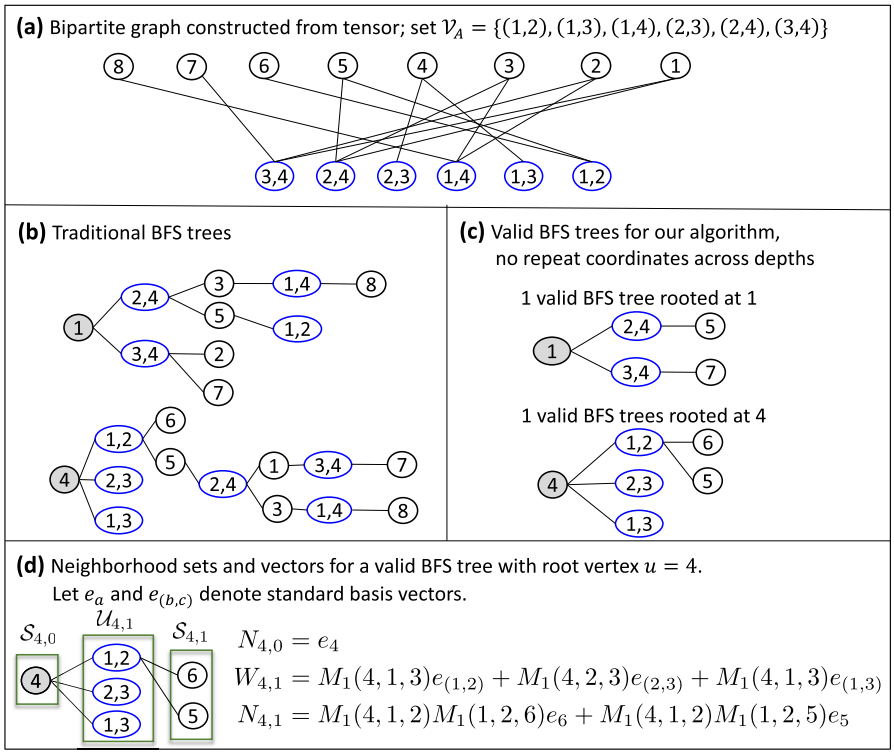

Fig. 3. Consider a symmetric 3-order tensor with $n = 8$, and the observation set $\Omega_1 = \{(1,2,4),(1,2,5),(1,2,6),(1,3,4),(1,4,8),(2,3,4),(2,4,5),(2,5,6),(3,4,7),(3,5,6)\}$. Figure **(a)** depicts the bipartite graph constructed from this set of observations. Weights would be assigned to edges based on the value of the observed entry in the tensor M_1. Figure **(b)** depicts the traditional notion of the BFS tree rooted at vertices 1 and 4. Vertices at layer/depth s correspond to vertices with shortest path distance of s to the root vertex. Figure **(c)** depicts valid BFS trees for our algorithm, which imposes an additional constraint that coordinates cannot be repeated across depths. For the BFS tree rooted at vertex 1, edges $((2,4),3)$ and $((3,4),2)$ are not valid, as coordinates 2 and 3 have both been visited in layer 2 by the vertices $(2,4)$ and $(3,4)$. For the BFS tree rooted at vertex 4, edge $(5,(2,4))$ is not valid as coordinate 2 has been visited in layer 2 by the vertex $(2,3)$ and coordinate 4 has been visited in layer 1 by the root vertex 4. Figure **(d)** depicts the sets $\mathcal{S}_{u,s}$ and $\mathcal{U}_{u,s}$ along with the neighborhood vectors $N_{u,s}$ and $W_{u,s}$ for a specific valid BFS tree rooted at vertex $u = 4$.

tensor, and each column corresponds to a pair of coordinates $(b,c) \in \mathcal{V}_A$ from the original tensor. We define M_B to be the n-by-$\binom{n/2}{2}$ matrix taking values $M_B(a,(b,c)) = M_1(a,b,c)$, where each row corresponds to an original coordinate of the tensor, and each column corresponds to a pair of coordinates $(b,c) \in \mathcal{V}_B$ from the original tensor. A row-column pair in the matrix corresponds to a triplet of coordinates in the original tensor. We will use matrices M_A and M_B to compute similarities or distances between coordinates, and we use tensor M_2 to compute the final estimates via nearest neighbor averaging. The data in M_A is used to construct depth $2t$ local neighborhoods rooted at each coordinate u, and M_B is used to connect the neighborhoods to form $2t+2$ length paths between any two coordinates u and a, which are then used to estimate the similarity between u and a. This approach is akin to the technique of "sprinkling" used in random graph analysis, in which we first analyze local neighborhoods formed with the edges in M_A, and then "sprinkle" the edges in M_B to connect these neighborhoods and argue that there are sufficiently many paths then that connect any two coordinates u and v.

Step 2: Construct Bipartite Graph From Ω_1, M_A: We define a bipartite graph corresponding to the flattened matrix M_A. Construct a graph with vertex set $[n] \cup \mathcal{V}_A$. There is an edge between vertex $a \in [n]$ and vertex $(b,c) \in \mathcal{V}_A$ if $(a,b,c) \in \Omega_1$, and the corresponding weight of the edge is $M_1(a,b,c)$. Recall that we assumed a symmetric model such that triplets that are permutations of one another will have the same data entry and thus the same edge weight in the associated graph. Figure 3(a) provides a concrete example of a bipartite graph constructed from tensor observations.

$$\begin{aligned}
&\mathbb{E}\left[M(u,e_1^1,e_2^1)\left(\prod_{i=1}^{t-1} M(e_1^i,e_2^i,x^i)M(x^i,e_1^{i+1},e_2^{i+1})\right) M(e_1^t,e_2^t,a) \;\middle|\; \theta_u,\theta_a \right] \\
&= \mathbb{E}\left[f(\theta_u,\theta_{e_1^1},\theta_{e_2^1})\left(\prod_{i=1}^{t-1} f(\theta_{e_1^i},\theta_{e_2^i},\theta_{x^i})f(\theta_{x^i},\theta_{e_1^{i+1}},\theta_{e_2^{i+1}})\right) f(\theta_{e_1^t},\theta_{e_2^t},\theta_a) \;\middle|\; \theta_u,\theta_a \right] \\
&= \sum_{k=1}^{r} \lambda_k^{2t} q_k(\theta_u) q_k(\theta_a) = e_u^T Q^T \Lambda^{2t} Q e_a. \qquad \text{(IV.2)}
\end{aligned}$$

Step 3: Expanding the Neighborhood: Consider the graph constructed from Ω_1, M_A. For each vertex $u \in [n]$, we construct a breadth first search (BFS) tree rooted at vertex u such that the vertices for each depth of the BFS tree consists only of new and previously unvisited coordinates, i.e. if vertex $a \in [n]$ is first visited at depth 4 of the BFS tree, then no vertex corresponding to (a, b) for any $b \in [n]$ can be visited in any subsequent depths greater than 4. Similarly, if (a', b') is visited in the BFS tree at depth 3, then vertices that include either of these coordinates, i.e. a', b', (a', c), or (b', c) for any $c \in [n]$, can not be visited in subsequent depths greater than 3. This restriction is only across different depths; we allow (a, b) and (a, c) to be visited at the same depth of the BFS tree.

There may be multiple valid BFS trees due to different ordering of visiting edges at the same depth. For example, if a vertex at depth s has edges to two different vertices at depth $s-1$ (i.e. two potential parents), only one of the edges can be chosen to maintain the tree property, but either choice is equally valid. Let us assume that when there is more than one option, one of the valid edges are chosen uniformly at random. Figure 3(c) shows valid BFS trees for a bipartite graph constructed from an example tensor.

The graph is bipartite so that each subsequent layer of the BFS tree alternates between the vertex sets $[n]$ and $\mathcal{V}_A$. Consider a valid BFS tree rooted at vertex $u \in [n]$ which respects the constraint that no coordinate is visited more than once. We will use $\mathcal{U}_{u,s} \subseteq \mathcal{V}_A$ to denote the set of vertices at depth $(2s-1)$ of the BFS tree, and we use $\mathcal{S}_{u,s} \subseteq [n]$ to denote the set of vertices at depth $2s$ of the BFS tree. Let $\mathcal{B}_{u,s} \subset [n] \cup \mathcal{V}_A$ denote the set of vertices which are visited in the first s layers of the BFS tree,

$$\mathcal{B}_{u,s} = \cup_{h \in \lfloor s/2 \rfloor} \mathcal{S}_{u,h} \cup_{l \in \lceil s/2 \rceil} \mathcal{U}_{u,l}.$$

We will overload notation and sometimes use $\mathcal{B}_{u,s}$ to denote the subset of coordinates in $[n]$ visited in the first s layers of the BFS tree, including both visited single coordinate vertices or coordinates in vertices $\mathcal{V}_A$, i.e.

$$\begin{aligned}\mathcal{B}_{u,s} =& \cup_{h \in \lfloor s/2 \rfloor} \mathcal{S}_{u,h} \\ &\cup \{x \in [n] \text{ s.t. } \exists (y,z) \in \cup_{l \in \lceil s/2 \rceil} \mathcal{U}_{u,l} \text{ satisfying } x \in \{y, z\}\}.\end{aligned}$$

Let $\mathcal{G}(\mathcal{B}_{u,s})$ denote all the information corresponding to the subgraph restricted to the first s layers of the BFS tree rooted at u. This includes the vertex set $\mathcal{B}_{u,s}$, the latent variables $\{\theta_a\}_{a \in \mathcal{B}_{u,s}}$ and the edge weights $\{M_1(a,b,c)\}_{a,(b,c) \in \mathcal{B}_{u,s}}$.

We define neighborhood vectors which represent the different layers of the BFS tree. Let $N_{u,s} \in [0,1]^n$ be associated to set $\mathcal{S}_{u,s}$, where the a-th coordinate is equal to the product of weights along the path from u to a in the BFS tree for $a \in \mathcal{S}_{u,s}$. Similarly, let $W_{u,s} \in [0,1]^{\mathcal{V}_A}$ be associated to set $\mathcal{U}_{u,s}$, where the (b,c)-th coordinate is equal to the product of weights along the path from u to (b,c) in the BFS tree for $(b,c) \in \mathcal{U}_{u,s}$. For $a \in [n]$, let $\pi_u(a)$ denote the parent of a in the valid BFS tree rooted at vertex u. For $(b,c) \in \mathcal{V}_A$, let $\pi_u(b,c)$ denote the parent of (b,c) in the BFS tree rooted at vertex u. We can define the neighborhood vectors recursively,

$$\begin{aligned}N_{u,s}(a) &= M_A(a, \pi_u(a)) W_{u,s}(\pi_u(a)) \mathbb{I}_{(a \in \mathcal{S}_{u,s})} \\ W_{u,s}(b,c) &= M_A(\pi_u(b,c), (b,c)) N_{u,s-1}(\pi_u(b,c)) \mathbb{I}_{((b,c) \in \mathcal{U}_{u,s})}\end{aligned}$$

and $N_{u,0} = e_u$. Let $\tilde{N}_{u,s}$ denote the normalized vector $\tilde{N}_{u,s} = N_{u,s}/|\mathcal{S}_{u,s}|$ and let $\tilde{W}_{u,s}$ denote the normalized vector $\tilde{W}_{u,s} = W_{u,s}/|\mathcal{U}_{u,s}|$. Figure 3(d) illustrates the neighborhood sets and vectors for a valid BFS tree. Recall from Eq. (IV.2), shown at the bottom of the previous page, that conditioned on θ_u and θ_a, $\mathbb{E}[N_{u,s}(a) \mid \theta_u, \theta_a] = \mathbb{P}(a \in \mathcal{S}_{us})\, e_u^T Q^T \Lambda^{2s} Q e_a$. Furthermore the event that $a \in \mathcal{S}_{us}$ only depends on the presence of the edges as determined by the Bernoulli uniform sampling, and is thus independent from the latent variable θ_a. We will show in a subsequent Lemma 7.2 that $e_k^T Q \tilde{N}_{u,t} \approx e_k^T \Lambda^{2t} Q e_u$, implying that the neighborhood vector $\tilde{N}_{u,t}$, which is constructed from products of weights over length $2s$ paths originating at u, is a statistic that is approximates $\Lambda^{2t} Q e_u$. Similar calculations show that $\mathbb{E}[W_{u,s}(b,c) \mid \theta_u, \theta_b, \theta_c] = \mathbb{P}((b,c) \in \mathcal{U}_{us}) \sum_{k=1}^r \lambda_k^{2t-1} q_k(\theta_u) q_k(\theta_b) q_k(\theta_c)$.

Step 4: Computing the Distances Using M_B: Let

$$t = \left\lceil \frac{\ln(n)}{2\ln(p^2 n^3)} \right\rceil. \tag{IV.3}$$

A heuristic for the distance would be

$$\begin{aligned}\mathsf{dist}(u,v) &\approx \frac{1}{|\mathcal{V}_B| p^2} (\tilde{N}_{u,t} - \tilde{N}_{v,t}) M_B M_B^T (\tilde{N}_{u,t} - \tilde{N}_{v,t}) \\ &= \frac{1}{|\mathcal{V}_B| p^2} \sum_{(\alpha,\beta) \in \mathcal{V}_B} \sum_{a,b \in [n]^2} (\tilde{N}_{u,t}(a) - \tilde{N}_{v,t}(a)) M_B(a,\alpha,\beta) \\ &\qquad \times M_B(b,\alpha,\beta)(\tilde{N}_{u,t}(b) - \tilde{N}_{v,t}(b))\end{aligned} \tag{IV.4}$$

For technical reasons that facilitate cleaner analysis, we use the following distance calculations. There are two deviations from the equation in (IV.4). First we exclude $a = b$ from the summation. Second we exclude coordinates for α or β that have been visited previously in $\mathcal{B}_{u,2t}$ or $\mathcal{B}_{v,2t}$. Define distance as

$$\mathsf{dist}(u,v) = (Z_{uu} + Z_{vv} - Z_{uv} - Z_{vu}), \tag{IV.5}$$

$$\begin{aligned}Z_{uv} &= \frac{1}{|\mathcal{V}_B(u,v,t)| p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{(\alpha,\beta) \in \mathcal{V}_B(u,v,t)} T_{uv}(\alpha,\beta), \\ \mathcal{V}_B(u,v,t) &= \{(\alpha,\beta) \in \mathcal{V}_B \text{ s.t. } \alpha \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}, \beta \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}\}, \\ T_{uv}(\alpha,\beta) &= \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) M_B(a,(\alpha,\beta)) M_B(b,(\alpha,\beta)).\end{aligned} \tag{IV.6}$$

We will show in Lemma 6.2 that $\mathsf{dist}(u,v) \approx d(u,v) := \|\Lambda^{2t+1} Q(e_u - e_v)\|_2^2$. The estimate is constructed by averaging over the product of weights over paths from u to v, where the term $N_{u,t}(a) N_{v,t}(b) M_B(a,(\alpha,\beta)) M_B(b,(\alpha,\beta))$ in Eq. (IV.6) is the product of weights over the path that goes from u to a to (α,β) to b to v, as $N_{u,t}(a)$ represents the products of weights on the path from u to a and $N_{v,t}(b)$ represents the products of weights on the path from b to v. The parameter t is chosen such that there are sufficiently many paths that we are averaging over in order to reduce the noise. In particular, the choice of $t \geq \ln(n)/2\ln(p^2 n^3)$ implies that $|\mathcal{S}_{u,t}| \geq (p^2 n^3)^t = \Omega(n^{1/2})$, such that the number of paths the estimator averages over is approximately $n^2 p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}| = \Omega(p^2 n^3)$. This rough calculation highlights that p must be $\omega(n^{-3/2})$ to guarantee that the number of paths used to compute $\mathsf{dist}(u,v)$ is increasing with n.

Step 5: Averaging Datapoints to Produce Final Estimate Let Ω_{2uvw} denote the set of indices (a,b,c) such that $a \leq b \leq c$, $(a,b,c) \in \Omega_2$, and the estimated distances $\mathsf{dist}(u,a)$, $\mathsf{dist}(v,b)$, $\mathsf{dist}(w,c)$ are all less than some chosen threshold

parameter η. The final estimate averages the datapoints corresponding to indices in Ω_{2uvw},

$$\hat{F}(u,v,w) = \frac{1}{|\Omega_{2uvw}|}\sum_{(a,b,c)\in\Omega_{2uvw}} M_2(a,b,c). \quad \text{(IV.7)}$$

B. Difference Between Tensor and Matrix Setting

The modifications in the construction of the breadth-first-search (BFS) tree for the tensor setting relative to the matrix setting are critical to the analysis. If we simply considered the classical construction of a BFS tree in the associated bipartite graph (as the matrix setting uses), this would lead to higher variance and bias due to the correlations of vertices sharing common latent variables associated to the same underlying coordinates of the tensor. Alternatively, if one constructed a BFS tree by not allowing any coordinate of the tensor to be visited more than once, this would also lead to suboptimal results as it would throw away too many entries, limiting the computed statistic to only order n data points. Our final algorithm, which allows for vertices with shared coordinates in the same depth of the BFS but not across different depths, is carefully chosen in order to break dependencies across different depths of the BFS tree, while still allowing for sufficient expansion in each depth.

To extend the algorithm to d-order tensors for $d > 3$, we will compute similarities between u, v via a similar computation as described in Steps 2-4 above, except it would be applied to the $n \times n^{d-1}$ matrix and associated bipartite graph corresponding to an unfolding of the tensor. Given the similarity estimates for pairwise coordinates, the final estimate would result from a standard nearest neighbor estimator over the high dimensional tensor. The primary part of the proof that would need to be modified is the analysis of the neighborhood vectors $N_{u,s}$ and W_{us} in step 3, which may involve constraining the growth of the BFS trees such that they extend deep enough before exhausting the visited coordinates.

V. Main Result

We provide an upper bound on the mean squared error (MSE) as well as the max entry-wise error (MEE) for the algorithm, showing that both the MSE and the MEE converge to zero as long as $p = n^{-3/2+\kappa}$ for some $\kappa > 0$. Our result implies that the simple variant of collaborative filtering algorithm based on estimating similarities produces a consistent estimator when the tensor latent function has finite spectrum or low rank. Further we show that it is robust to arbitrary, additive perturbation in that the estimation error increases gracefully in the amount of perturbation. To the best of our knowledge, such robustness to arbitrary bounded additive noise with respect to max-norm estimation is first of its kind in the literature on tensor estimation.

A. Finite Spectrum

We establish consistency of our estimator with respect to MSE and max-norm error of the algorithm when the underlying f has finite spectrum, i.e. rank r model with $r = \Theta(1)$.

Theorem 5.1: We assume that the function f is rank r, L-Lipschitz and that $\theta \sim U[0,1]$. Assume that $p = n^{-3/2+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$. Let t be defined according to (IV.3). For any arbitrarily small $\psi \in (0, \min(\kappa, \frac{3}{8}))$, choose the threshold

$$\eta = \Theta\left(n^{-(\kappa-\psi)}\right).$$

The algorithm produces estimates so that,

$$\text{MSE} = O(n^{-(\kappa-\psi)}) = O\left(\frac{n^{\psi}}{(p^2n^3)^{1/2}}\right),$$

and

$$\|F - \hat{F}\|_{\max} = O(n^{-(\kappa-\psi)/2}),$$

with probability $1 - O\left(n^4 \exp(-\Theta(n^{2\psi}))\right)$.

The error bounds in Theorem 5.1 imply that our estimator is consistent as long as $p = n^{-3/2+\kappa}$ for some $\kappa > 0$, with a MSE that scales as $O(1/pn^{3/2})$. The threshold of $p = \Omega(n^{-3/2})$ is optimal, and furthermore this requirement is precisely the threshold at which the constructed bipartite graph in Step 2 of the algorithm is fully connected. Below the connectivity threshold, the graph will be disconnected into small components with insufficient information to recover the expected value of edges across disconnected components.

In comparison to the literature, [31] prove that the minimax optimal MSE is $O(1/pn^2)$, which is achieved via spectral initialization followed by power iteration [31] or gradient descent [40] as long as $p = \Omega(n^{-3/2})$. While our result achieves a similar sample complexity threshold, our MSE rate is suboptimal by a factor of $\sqrt{n}$. A limitation of neighborhood smoothing is that we do not achieve exact recovery under the noiseless setting, and we do not achieve the minimax optimal rates. It is unclear whether the gap is due to a limitation in the analysis or the algorithm. A benefit of our analysis in contrast to the literature is that the neighborhood smoothing approach can more easily deal with approximate low rank settings as arise under smoothness, as presented in Theorem 5.2. While low rank is often a useful modeling concept for real world datasets, in reality most real-world applications are likely only approximately low rank rather than exactly low rank.

B. Approximately Finite Spectrum

For approximate rank r model, we establish a natural perturbation result for the algorithm. Specifically, if the underlying model has ε-approximate rank r, then we argue that the result of Theorem 5.1 remain true, both with respect to MSE and max-norm error, with perturbation amount of $\mathsf{poly}(\varepsilon)$.

Theorem 5.2: We assume that the function f has ε-approximate rank r, L-Lipschitz and that $\theta \sim U[0,1]$. Assume that $p = n^{-3/2+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$. Choosing t according to (IV.3), it follows that $t = \lceil \frac{1}{4\kappa} \rceil$. For any arbitrarily small $\psi \in (0, \min(\kappa, \frac{3}{8}))$, choose the threshold

$$\eta = \Theta\left(n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right).$$

The algorithm produces estimates so that,

$$\begin{aligned}\text{MSE} &= O(n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}) \\ &= O\left(\frac{n^{\psi}}{(p^2n^3)^{1/2}} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right),\end{aligned}$$

and

$$\|F - \hat{F}\|_{\max} = O(n^{-(\kappa-\psi)/2} + t\varepsilon(1+\varepsilon)^{2t-1} + \sqrt{t\varepsilon(1+\varepsilon)^{2t-1}}),$$

with probability $1 - O\left(n^4 \exp(-\Theta(n^{2\psi}))\right) - O(n^{-2})$.

As the entries of F are normalized such that $\|F\|_{\max} \leq 1$, the bound is meaningful when $\varepsilon < 1$, in which case the dominating term of the additional error due to the perturbation is linear in ε, as t is a constant. The proof of Theorem 5.2 relies on the following observation: the distribution of the data under the setting where the latent function f has ε-rank r is equivalent to the distribution of data generated according to the rank r approximation of f and then adding a deterministic perturbation to each observation accounting for the difference between f and its rank r approximation f_r, which is entrywise bounded by ε. In particular, the proof of Theorem 5.2 shows that under arbitrary deterministic perturbation of a rank r model where the perturbation is bounded by ε, the estimation error is perturbed by at most $\text{poly}(\varepsilon)$. As a byproduct, our result proves that our estimator is robust to arbitrary deterministic bounded noise in the observations.

The approximation guarantee depends on the spectral decay. Since the analysis allows any arbitrary adversarial model for the ε deviation from a low rank model, the resulting guarantee depending polynomially in ε is qualitatively best one can hope for. By definition, the minimal error must be lower bounded by ε, but determining the best achievable error as a function of ε is an important open question for future investigation. It is worth noting that no other prior work addresses such a robust error model.

C. Reducing Computational Complexity

The computational complexity can be estimated by analyzing steps 3-5 of the algorithm. Step 3 costs $O(pn^4)$, as there are n BFS trees to construct, which each take at most pn^3 edge traversals as there are at most order pn^3 edges in the constructed graph. Step 4 costs $O(p^2n^6)$ as there are order n^2 pairwise distances to compute, and each computed distance involves sums over terms indexed by $a, b, \alpha, \beta \in [n]^4$ where (a, α, β) and (b, α, β) are in the observation set. As the sparsity of the dataset is p, this results in order p^2n^4 nonzero terms in the summation, each of which is the product of 4 quantities, taking $O(1)$ to compute. Step 5 costs $O(pn^6)$ as there are $\Theta(n^3)$ triplets we need to estimate, and each involves averaging at most $O(pn^3)$ datapoints. In summary, the computation cost of the entire method, for $p = n^{-\frac{3}{2}+\kappa}$ is $O(pn^4 + p^2n^6 + pn^6)$ where the cost in Step 5 dominates.

This computation cost can be improved drastically. For example, as explained in [18], by use of 'representative' or 'anchor' vertices chosen as random, the algorithm can instead cluster the vertices with respect to these anchor vertices and learn a block constant estimate, significantly reducing the involved computation. If there are y anchor vertices, then Step 4 reduces to only computing pairwise distances between $\binom{y}{2} + ny$ pairs of vertices, as non-anchor vertices are only compared to the small set of y anchor vertices. Step 5 reduces to only estimating $\binom{y}{3}$ entries of the tensor corresponding to combinations of the anchor vertices, and then extrapolating the estimate to other vertices assigned to the same cluster. This would result in a computational cost of $O(pn^4 + (y^2 + ny)p^2n^4 + y^3pn^3)$. When $p = n^{-\frac{3}{2}+\kappa}$, our proof indicates that by choosing $y = \Theta((p^2n^3)^{1/4}) = \Theta(n^{\kappa/2})$, the corresponding block constant estimator would achieve the same rates on the MSE and MEE as presented in Theorem 5.1, while requiring a reduced computational complexity of $O(n^{5/2+\kappa} + n^{2+5\kappa/2})$.

Corollary 5.3: We assume that the function f is rank r, L-Lipschitz and that $\theta \sim U[0,1]$. Assume that $p = n^{-3/2+\kappa}$ for some $\kappa \in (0, \frac{1}{2})$. Let t be defined as per (IV.3). For any arbitrarily small $\psi \in (0, \min(\kappa, \frac{3}{8}))$, choose the threshold

$$\eta = \Theta\left(n^{-(\kappa-\psi)}\right).$$

The modified algorithm which subsamples $y = \Omega((p^2n^3)^{1/4}) = \Omega(n^{\kappa/2})$ anchor vertices at random and uses them to cluster the vertices to learn a block constant estimate will achieve

$$\text{MSE} = O(n^{-(\kappa-\psi)}) = O\left(\frac{n^{\psi}}{(p^2n^3)^{1/2}}\right),$$

and

$$\|F - \hat{F}\|_{\max} = O(n^{-(\kappa-\psi)/2}),$$

with probability $1 - O\left(n^4 \exp(-\Theta(n^{2\psi}))\right)$.

D. Discussion of Assumptions

We assumed in our algorithm and analysis that we had two fresh samples of the dataset, M_1 and M_2. The dataset M_1 is used to estimate distances between coordinates, and the dataset M_2 is used to compute the final nearest neighbor estimates. Given only a single dataset, the same theoretical results can also be shown by simply splitting the samples uniformly into two sets, one used to estimate distances and one used to compute the nearest neighbor estimates. as we are considering the sparse regime with $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$, the two subsets after sample splitting will be nearly independent, such that the analysis only needs to be slightly modified. This is formally handled in the paper on collaborative filtering for matrix estimation in [18].

Our model and analysis assumes that the latent variables $\{\theta_u\}_{u\in[n]}$ are sampled uniformly on the unit interval, and that the function f is Lipschitz with respect to θ. This assumption can in fact be relaxed significantly, as it is only used in the final step of the proof in analyzing the nearest neighbor estimator. Proving that the distance estimates concentrate well does not require these assumptions, in particular it primarily uses the low rank assumption. Given that the distance estimate concentrates well, the analysis of the nearest neighbor estimator depends on the local measure, i.e. what fraction of other coordinates have similar function values so that the estimated distance is small. We used Lipschitzness and uniform distribution on the unit interval in order to lower bound the fraction of nearby coordinates, however many other properties would also lead to such a bound. The dependence of the noisy nearest neighbor estimator on the local measure is discussed in detail in [16]. Similar extensions as presented in [16] would apply

for our analysis here, leading to consistency and convergence rate bounds for examples including when

- the latent space has only finitely many elements, or equivalently the distribution of θ has finite support;
- the latent space is the unit hypercube in a finite dimensional space and the latent function is Lipschitz;
- the latent space is a complete, separable metric space, i.e. Polish space, with bounded diameter and the latent function is Lipschitz.

Although our stated results assume a symmetric tensor, the results naturally extend to asymmetric $(n_1 \times n_2 \times n_3)$ tensors as long as n_1, n_2, and n_3 are proportional to one another. Our analysis can be modified for the asymmetric setting, or one can reduce the asymmetric tensor to a $(n \times n \times n)$ symmetric tensor where $n = n_1 + n_2 + n_3$, and the coordinates of the new tensor consists of the union of the coordinates in all three dimensions of the asymmetric tensor. The results applied to this larger tensor would still hold with adjustments of the model allowing for piecewise Lipschitz functions.

In the proof sketch that follows below, we show that for the 3-order tensor, the sample complexity threshold of $p = \omega(n^{-3/2})$ directly equals the density of observations needed to guarantee the bipartite graph is connected with high probability. Although our stated results assume a 3-order tensor, our algorithm and analysis can be likely extended to general d-order tensors. The proof would instead require analysis of the $n \times n^{d-1}$ matrix and associated bipartite graph corresponding to the unfolding of the tensor. The bipartite graph would consist of vertex sets $[n]$ and $[n]^{d-1}$.

Remark 5.1: In order for the vertices in $[n]$ to be fully connected to each other, p needs to be $\Omega(n^{d/2})$, which can be proved using the standard branching process analysis as is used to prove the connectivity threshold of an Erdos Renyi graph [45]. Let X_u denote the set of vertices $v \in [n]$ such that the distance between u and v in the bipartite graph is 2. It follows that $\mathbb{P}(v \in X_u) = 1 - (1-p^2)^{n^{d-1}}$, such that $\mathbb{E}[|X_u|] = (n-1)(1-(1-p^2)^{n^{d-1}}) = \Theta(p^2 n^d)$. As a result, if $p = o(n^{d/2})$, it follows that $\mathbb{P}(X_u = \emptyset) = 1 - \mathbb{P}(|X_v| \geq 1) \geq 1 - \mathbb{E}[|X_u|] = 1 - o(1)$, such that with probability tending to 1 the vertex u will be isolated, i.e. not connected to any other vertex $v \in [n]$. To prove that the graph is connected for $p = \tilde{\Omega}(n^{d/2})$, one would relate the growth of a local neighborhood in the graph to an appropriate branching process, where the expected number of descendents alternates between pn and pn^{d-1}. To formalize the argument, one would then argue that the branching process survives to infinity if $p^2 n^d$ is sufficiently large, e.g polylog(n), and also argue that the branching process is a reasonable approximation for the neighborhood growth of the graph until a linear number of vertices are visited. The requirement for graph connectivity in our algorithm and analysis arises from the similarity computation $\mathsf{dist}(u, v)$, which involves products of weights over paths in the graph that connect u and v. The fact that $n^{d/2}$ also corresponds to the connectivity threshold in the corresponding bipartite graph sheds light on the computational lower bound for tensor completion in [27], giving another way to explain the $\Omega(n^{d/2})$ lower bound.

VI. Proof

In this section, we present the proof for Theorem 5.1. The proof outline is similar to the matrix setting in [18], in that the core of the analysis is proving that the distance function as defined in (IV.5) concentrates appropriately and captures an appropriate notion of distance that enables the classical "nearest neighbor" algorithm to be effective. However, due to high-dependencies across latent factors associated with columns that share tensor coordinates, the concentration of the BFS neighborhood expansion in section VII-B requires a new argument beyond the simple martingale argument in the matrix setting. This involves a careful application of the concentration of U-statistics. Furthermore, the concentration of the distance calculation in Eq (IV.6) as analyzed in section VII-D requires a new argument relating the computed statistic to a thresholded variant more amenable to analysis. This is due to both the dependencies in the latent factors along with the lopsidedness in the dimensions so that straightforward applications of standard concentration results are too weak and insufficient to drive the error to zero.

While the proof is stated for bounded observations, i.e. bounded noise, the result can be extended to sub-Gaussian noise rather than uniformly bounded noise. This would involve showing that the norms of the neighborhood vectors N_{us} and W_{us} are well-controlled such that the application of Hoeffding's inequality used in Lemmas 7.4 and 7.5 still hold. Additionally the proof of 7.7 naively bounds the products of weights over a path in absolute value by 1; if the noise were not bounded but sub-Gaussian, one would have to additionally argue that the product of the weights would be sufficiently controlled with high probability.

The critical lemma that the proof hinges on shows that the computed similarities, i.e. $\mathsf{dist}(u, a)$, concentrates around the function $d(u, a) = \|\Lambda^{2t+1} Q(e_u - e_a)\|_2^2$. This then implies that if $d(u, a)$, $d(v, b)$ and $d(w, c)$ are small, the function value $F(u, v, w)$ would be close to $F(a, b, c)$. Additionally, we use Lipschitzness of the latent function f along with the assumption that θ_u are sampled independently from $U[0, 1]$ to argue that for any u, there is a sufficiently large set of other coordinates a such that $d(u, a)$ is small. If these properties hold, then a simple analysis of nearest neighbor averaging using $\mathsf{dist}(u, a)$ to determine the neighbors will result in a bias variance tradeoff that can be tuned to show our final results. As such, the complexity of the proof revolves around showing that the computed $\mathsf{dist}(u, a)$ concentrates around $d(u, a)$. This involves a delicate analysis which involves first arguing that the normalized neighborhood vectors $\tilde{N}_{u,t}$ satisfy $e_k^T Q \tilde{N}_{u,t} \approx e_k^T \Lambda^{2t} Q e_u$, which involves martingale concentration as well as concentration of appropriately defined U-statistics. Subsequently we need to argue that the statistic

$$Z_{uv} = \frac{1}{|\mathcal{V}_B(u, v, t)| p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \times \sum_{(\alpha,\beta) \in \mathcal{V}_B(u,v,t)} \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) M_B(a, (\alpha, \beta)) M_B(b, (\alpha, \beta)) \approx \tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t}.$$

A challenge in the analysis is that each term in the sum has very small probability of being nonzero such that the sum is

sparse enough that the standard concentration inequalities are not tight enough. Thus we relate the sum to a thresholded variant and use a tighter approximation of the binomial cdf to obtain the desired bound.

A. Analyzing Noisy Nearest Neighbors

We start by stating an important Lemma 6.1, adapted from [18] that characterizes the error of the noisy nearest neighbor algorithm. Recall that our algorithm estimates $F(u,v,w)$, i.e. $f(\theta_u,\theta_v,\theta_w)$, according to (IV.7), which simply averages over data-points $M_2(a,b,c)$ corresponding to tuples (a,b,c) for which a is close to u, b is close to v and c is close to w according to the estimated distance function. The choice of parameter η allows for tradeoff between bias and variance of the algorithm.

We first argue that the data-driven distance estimates dist will concentrate around an ideal data-independent distance $d(\theta_u,\theta_v)$ for $d:[0,1]^2\to\mathbb{R}_+$. We subsequently argue that the nearest neighbor estimate produced by (IV.7) using $d(\theta_u,\theta_v)$ in place of $\mathsf{dist}(u,v)$ will yield a good estimate by properly choosing the threshold η to tradeoff between bias and variance. The bias will depend on the local geometry of the function f relative to the distances defined by d. The variance depends on the measure of the latent variables $\{\theta_u\}_{u\in[n]}$ relative to the distances defined by d, i.e. the number of observed tuples $(a,b,c)\in\Omega_2$ such that $d(\theta_u,\theta_a)\le\eta$, $d(\theta_v,\theta_b)\le\eta$ and $d(\theta_w,\theta_c)\le\eta$ needs to be sufficiently large. We formalize the above stated desired properties.

Property 6.1 (Good Distance): We call an ideal distance function $d:[0,1]^2\to\mathbb{R}_+$ to be a bias-good distance function for some $\mathsf{bias}:\mathbb{R}_+\to\mathbb{R}_+$ if for any given $\eta>0$ it follows that $|f(\theta_a,\theta_b,\theta_c)-f(\theta_u,\theta_v,\theta_w)|\le\mathsf{bias}(\eta)$ for all $(\theta_a,\theta_b,\theta_c,\theta_u,\theta_v,\theta_w)\in[0,1]^4$ such that $d(\theta_u,\theta_a)\le\eta$, $d(\theta_v,\theta_b)\le\eta$ and $d(\theta_w,\theta_c)\le\eta$.

Property 6.2 (Good Distance Estimation): For some $\Delta>0$, we call distance $\hat{d}:[n]^2\to\mathbb{R}_+$ a Δ-good estimate for ideal distance $d:[0,1]^2\to\mathbb{R}_+$, if $|d(\theta_u,\theta_a)-\hat{d}(u,a)|\le\Delta$ for all $(u,a)\in[n]^2$.

Property 6.3 (Sufficient Representation): The collection of coordinate latent variables $\{\theta_u\}_{u\in[n]}$ is called meas-represented for some $\mathsf{meas}:\mathbb{R}_+\to\mathbb{R}_+$ if for any $u\in[n]$ and $\eta'>0$, $\frac{1}{n}\sum_{a\in[n]}\mathbb{I}_{(d(u,a)\le\eta')}\ge\mathsf{meas}(\eta')$.

Lemma 6.1: Assume that property 6.1 holds with probability 1, property 6.2 holds for any given pair $u,a\in[n]$ with probability $1-\alpha_1$, and property 6.3 holds with probability $1-\alpha_2$ for some η,Δ, and $\eta'=\eta-\Delta$; in particular d is a bias-good distance function, $\hat{d}=\mathsf{dist}$ as estimated from M_A is a Δ-good distance estimate for d, and $\{\theta_u\}_{u\in[n]}$ is meas-represented. Then noisy nearest neighbor estimate $\hat{F}$ computed according to (IV.7) satisfies

$$\mathrm{MSE}(\hat{F})\le\mathsf{bias}^2(\eta+\Delta)+\frac{\sigma^2}{(1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3}+\exp\left(-\frac{\delta^2p\left(\mathsf{meas}(\eta-\Delta)n\right)^3}{2}\right)+3\ n\alpha_1+\alpha_2,$$

for any $\delta\in(0,1)$. Furthermore, for any $\delta'\in(0,1)$ and $(u,v,w)\in[n]^3$,

$$|\hat{F}(u,v,w)-f(\theta_u,\theta_v,\theta_w)|\le\mathsf{bias}(\eta+\Delta)+\delta',$$

with probability at least

$$1-\exp\left(-\tfrac{1}{2}\delta^2p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right)-\exp\left(-\delta'^2(1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right)-3\ n\alpha_1-\alpha_2.$$

The proof of Lemma 6.1 is a modification from [18] and is included in the Appendix.

B. Proofs of Theorems 5.1 and 5.2

Proof: We prove that as long as $p=n^{-3/2+\kappa}$ for any $\kappa\in(0,\frac{1}{2})$, with high probability, properties 6.1-6.3 hold for an appropriately chosen function d, and for distance estimates $\hat{d}=\mathsf{dist}$ computed according to (IV.5) with t defined in (IV.3). We subsequently use Lemma 6.1 to conclude Theorem 5.1 and Theorem 5.2. The proofs of Properties 6.1 and 6.3 are identical in Theorem 5.1 and Theorem 5.2, while that of property 6.2 differ. For Theorem 5.1, we utilize Lemma 6.2 while for Theorem 5.2, we utilize Lemma 6.3. The proof of Theorem 5.2 follows nearly the same argument, where f will be replaced by the rank r approximation f_r, c.f. (III.4).

Good Distance d and Property 6.1: We start by defining the ideal distance d as follows. For all $(u,v)\in[n]^2$, let

$$\begin{aligned}d(\theta_u,\theta_v)&=\|\Lambda^{2t+1}Q(e_u-e_v)\|_2^2\\&=\sum_{k=1}^{r}\lambda_k^{2(2t+1)}(q_k(\theta_u)-q_k(\theta_v))^2.\end{aligned}\tag{VI.1}$$

Recall that t is defined in (IV.3). Since $p=n^{-3/2+\kappa}$ and $\kappa\in(0,\frac{1}{2})$, we have that

$$t=\left\lceil\frac{\ln(n)}{2\ln(p^2n^3)}\right\rceil=\left\lceil\frac{1}{4\kappa}\right\rceil.\tag{VI.2}$$

We want to show that there exists $\mathsf{bias}:\mathbb{R}_+\to\mathbb{R}_+$ so that $|(f(\theta_a,\theta_b,\theta_c)-f(\theta_u,\theta_v,\theta_w))|\le\mathsf{bias}(\eta)$ for any $\eta>0$ and $(u,a,v,b,w,c)\in[n]^3$ such that $d(\theta_u,\theta_a)\le\eta$, $d(\theta_v,\theta_b)\le\eta$ and $d(\theta_w,\theta_c)\le\eta$. Consider

$$\begin{aligned}&|f(\theta_u,\theta_v,\theta_w)-f(\theta_a,\theta_b,\theta_c)|\\&\le|f(\theta_u,\theta_v,\theta_w)-f(\theta_a,\theta_v,\theta_w)|\\&\quad+|f(\theta_a,\theta_v,\theta_w)-f(\theta_a,\theta_b,\theta_w)|\\&\quad+|f(\theta_a,\theta_b,\theta_w)-f(\theta_a,\theta_b,\theta_c)|.\end{aligned}\tag{VI.3}$$

Now

$$\begin{aligned}&|f(\theta_u,\theta_v,\theta_w)-f(\theta_a,\theta_v,\theta_w)|\\&=|\sum_k\lambda_k(q_k(\theta_u)-q_k(\theta_a))q_k(\theta_v)q_k(\theta_w)|\\&\overset{(a)}{\le}B^2|\sum_k\lambda_k(q_k(\theta_u)-q_k(\theta_a))|\\&=B^2\|\Lambda Q(e_u-e_a)\|_1\\&\le B^2\sqrt{r}\|\Lambda Q(e_u-e_a)\|_2\\&\le B^2\sqrt{r}|\lambda_r|^{-2t}\|\Lambda^{2t+1}Q(e_u-e_a)\|_2\\&=B^2\sqrt{r}|\lambda_r|^{-2t}\sqrt{d(\theta_u,\theta_a)}.\end{aligned}\tag{VI.4}$$

In above, (a) follows from the $\|q_k(\cdot)\|_\infty \le B$ for all k. Repeating this argument to bound the other terms in (VI.3), we obtain that

$$\begin{aligned}&|f(\theta_u,\theta_v,\theta_w) - f(\theta_a,\theta_b,\theta_c)|\\ &\le 3B^2\sqrt{r}|\lambda_r|^{-2t}\max\left(\sqrt{d(\theta_u,\theta_a)},\sqrt{d(\theta_v,\theta_b)},\sqrt{d(\theta_w,\theta_c)}\right)\\ &\le 3B^2|\lambda_r|^{-2t}\sqrt{r\eta} \;\equiv\; \mathsf{bias}(\eta). \end{aligned} \tag{VI.5}$$

In summary, property 6.1 is satisfied for distance function d defined according to (VI.1) and $\mathsf{bias}(\eta) = 3B^2|\lambda_r|^{-2t}\sqrt{r\eta}$.

Good Distance Estimate $\hat{d}$ and Property 6.2: We state the following Lemma whose proof is delegated to Section VII.

Lemma 6.2: Given f with rank r, assume that $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$. Let $\hat{d} = \mathsf{dist}$ as defined in (IV.5). Then for any $(u,a) \in [n]^2$, for any $\psi \in (0,\kappa)$,

$$|d(\theta_u,\theta_a) - \hat{d}(u,a)| = O\left(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)}\right),$$

with probability at least $1 - O\Big(\exp(-n^{2\psi}(1-o(1)))\Big)$. Lemma 6.2 implies that property 6.2 holds with probability $1-o(1)$ for $\Delta = \Theta\left(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)}\right)$ when f has rank r.

Lemma 6.3: Given f with ε-approximate rank r for $\varepsilon \ge 0$, assume that $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$. Let $\hat{d} = \mathsf{dist}$ as defined in (IV.5). Then for any $(u,a) \in [n]^2$, for any $\psi \in (0,\kappa)$,

$$\begin{aligned}|d(\theta_u,\theta_a) - \hat{d}(u,a)| &= O\left(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)}\right)\\ &\quad + O\left(t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right),\end{aligned}$$

with probability at least $1 - O\Big(\exp(-n^{2\psi}(1-o(1)))\Big) - O\Big(n^{-6}\Big)$. Lemma 6.3 implies that property 6.2 holds with probability $1-o(1)$ for

$$\Delta = \Theta\left(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right),$$

when f has ε-approximate rank r.

Sufficient Representation and Property 6.3: Since f is L-Lipschitz, the distance d as defined in (VI.1) is bounded above by the squared ℓ_2 distance:

$$\begin{aligned}&d(\theta_u,\theta_v)\\ &= \|\Lambda^{2t+1}Q(e_u - e_v)\|_2^2\\ &\le |\lambda_1|^{4t}\|\Lambda Q(e_u - e_v)\|_2^2\\ &= |\lambda_1|^{4t}\Big(\sum_{k=1}^r \lambda_k^2 (q_k(\theta_u) - q_k(\theta_v))^2\\ &= |\lambda_1|^{4t}\sum_{k=1}^r \lambda_k^2 (q_k(\theta_u) - q_k(\theta_v))^2\Big(\int_0^1 q_k(\theta_a)^2 d\theta_a\Big)\\ &\qquad\qquad \times \Big(\int_0^1 q_k(\theta_b)^2 d\theta_b\Big)\\ &= |\lambda_1|^{4t}\sum_{k=1}^r \lambda_k^2 \int_{[0,1]^2} (q_k(\theta_u)q_k(\theta_a)q_k(\theta_b)\\ &\qquad\qquad - q_k(\theta_v)q_k(\theta_a)q_k(\theta_b))^2 d\theta_a d\theta_b\Big)\\ &\overset{(a)}{=} |\lambda_1|^{4t}\int_0^1\int_0^1 (f(\theta_u,\theta_a,\theta_b) - f(\theta_v,\theta_a,\theta_b))^2 d\theta_a d\theta_b\\ &\le |\lambda_1|^{4t}L^2|\theta_u - \theta_v|^2, \end{aligned} \tag{VI.7}$$

where in (a) we have used the fact that $q_k(\cdot), k \in [r]$ are orthonormal with respect to uniform distribution over $[0,1]$. We assumed that the latent parameters $\{\theta_u\}_{u\in[n]}$ are sampled i.i.d. uniformly over $[0,1]$. Therefore, for any $\theta_u \in [0,1]$, for any $v \in [n]$ and $\eta' > 0$,

$$\begin{aligned}\mathbb{P}\left(d(\theta_u,\theta_v) \le \eta' \;\middle|\; \theta_u\right) &\ge \mathbb{P}\left(|\lambda_1|^{4t}L^2|\theta_u - \theta_v|^2 \le \eta' \;\middle|\; \theta_u\right)\\ &= \mathbb{P}\left(|\theta_u - \theta_v| \le \frac{\sqrt{\eta'}}{|\lambda_1|^{2t}L} \;\middle|\; \theta_u\right)\\ &\ge \min\left(1, \frac{\sqrt{\eta'}}{|\lambda_1|^{2t}L}\right). \end{aligned} \tag{VI.8}$$

Let us define

$$\mathsf{meas}(\eta') = \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^t L} \tag{VI.9}$$

for all $\eta' \in (0, |\lambda_1|^{4t}L^2)$. By an application of Chernoff's bound and a simple majorization argument, it follows that for all $\eta' \in (0, |\lambda_1|^{4t}L^2)$ and $\delta \in (0,1)$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n-1}\sum_{a\in[n]\setminus u} \mathbb{I}_{(d(u,a)\le\eta')} \le \mathsf{meas}(\eta') \;\middle|\; \theta_u\right)\\ &\qquad \le \exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^{2t}L}\right).\end{aligned}$$

By using union bound over all n indices, it follows that for any $\eta' \in (0, |\lambda_1|^{4t}L^2)$, with probability at least $1 - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^{2t}L}\right)$, property 6.3 is satisfied with meas as defined in (VI.9).

Concluding Proof of Theorem 5.1: In summary, property 6.1 holds with probability 1, by Lemma 6.2 property 6.2 holds for a given tuple $(u,a) \in [n]^2$ with probability $1-\alpha_1$ where $\alpha_1 = O\Big(\exp(-n^{2\psi}(1-o(1)))\Big)$ for $\psi \in (0, \min(\kappa, \frac{3}{8}))$ and $\kappa \in (0, \frac{1}{2})$, property 6.3 holds with probability $1-\alpha_2$ where $\alpha_2 = n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^{2t}L}\right)$ with distance estimate $\hat{d} = \mathsf{dist}$ defined in (IV.5) with

$$\begin{aligned} d(\theta_u,\theta_v) &= \|\Lambda^{2t+1}Q(e_u - e_v)\|_2^2,\\ \mathsf{bias}(\eta) &= 3B^2|\lambda_r|^{-2t}\sqrt{r\eta},\\ \Delta &= \Theta(r\lambda_{\max}^{4t}n^{-(\kappa-\psi)}),\\ \mathsf{meas}(\eta') &= \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^{2t}L}, \end{aligned} \tag{VI.10}$$

for any $\eta > 0$, $\delta \in (0,1)$ and $\eta' = \eta - \Delta \in (0, |\lambda_1|^{4t}L^2)$. By substituting the expressions for bias, meas, and α into Lemma 6.1, it follows that

$$\begin{aligned}&\mathrm{MSE}(\hat{F})\\ &\le 9B^4|\lambda_r|^{-4t}r(\eta+\Delta) + \frac{\sigma^2L^3|\lambda_1|^{6t}}{(1-\delta)^4 p\left(\sqrt{\eta-\Delta}n\right)^3}\\ &\quad + \exp\left(-\frac{\delta^2(1-\delta)^3 p\left(\sqrt{\eta-\Delta}n\right)^3}{2L^3|\lambda_1|^{6t}}\right)\\ &\quad + nO\Big(\exp(-n^{2\psi}(1-o(1)))\Big)\\ &\quad + n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^{2t}L}\right).\end{aligned}$$

Additionally, for any $\delta' \in (0,1)$,

$$|\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w)| \le 3B^2|\lambda_r|^{-2t}\sqrt{r(\eta+\Delta)} + \delta' \quad \text{(VI.11)}$$

with probability at least

$$1 - \exp\left(-\frac{\delta^2(1-\delta)^3 p\left(\sqrt{\eta-\Delta}n\right)^3}{2L^3|\lambda_1|^{6t}}\right) - \exp\left(-\frac{\delta'^2(1-\delta)^4 p\left(\sqrt{\eta-\Delta}n\right)^3}{L^3|\lambda_1|^{6t}}\right) - nO\Big(\exp(-n^{2\psi}(1-o(1)))\Big) - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^{2t}L}\right).$$

By selecting $\eta = \Theta(\Delta) = \Theta(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)})$ with a large enough constant so that $\eta - \Delta = \Theta(\eta)$, it follows that by the conditions that $\psi > 0$ and $\kappa < \frac{1}{2}$,

$$\eta \pm \Delta = \Theta(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)}),$$
$$p(\sqrt{\eta-\Delta}n)^3 = \Theta(r^{3/2}\lambda_{\max}^{6t} n^{\frac{3}{2}-\frac{\kappa}{2}+\frac{3\psi}{2}}) = \Omega(n^{\frac{5}{4}}),$$
$$n\sqrt{\eta-\Delta} = \Theta(r^{1/2}\lambda_{\max}^{2t} n^{1-\frac{\kappa-\psi}{2}}) = \Omega(n^{\frac{3}{4}}).$$

By substituting this choice of η and $\delta = \frac{1}{2}$, it follows that

$$\text{MSE}(\hat{F}) = O\Big(r^2(\lambda_{\max}/\lambda_r)^{4t} n^{-(\kappa-\psi)}\Big). \quad \text{(VI.12)}$$

By choosing $\delta' = n^{-(\kappa-\psi)/2}$ such that $\delta' = \Theta(\sqrt{\eta})$ and $\delta'^2 p(\sqrt{\eta-\Delta}n)^3 = \Omega(n^{\frac{3}{4}}) = \Omega(n^{2\psi})$ because $\psi < \frac{3}{8}$. Therefore, by substituting into (VI.11), it follows that for any given $(u,v,w) \in [n]^3$, with probability $1 - O(n\exp(-\Theta(n^{2\psi})))\Big)$,

$$|\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w)| = O\Big(r(\lambda_{\max}/\lambda_r)^{2t} n^{-(\kappa-\psi)/2}\Big). \quad \text{(VI.13)}$$

Using union bound over choices of $(u,v,w) \in [n]^3$, it follows that the maximum entry-wise error is bounded above by $O\Big(n^{-(\kappa-\psi)/2}\Big)$ with probability $1 - O(n^4\exp(-\Theta(n^{2\psi})))\Big)$. This completes the proof of Theorem 5.1. □

Concluding Proof of Theorem 5.2: We follow similar line of argument as for proof of Theorem 5.1. As noted earlier, property 6.1 holds with probability 1, by Lemma 6.3 property 6.2 holds for a given tuple $(u,a) \in [n]^2$ with probability $1-\alpha_1$ where $\alpha_1 = O\Big(\exp(-n^{2\psi}(1-o(1))) + n^{-6}\Big)$ for $\psi \in (0, \min(\kappa, \frac{3}{8}))$ and $\kappa \in (0, \frac{1}{2})$, property 6.3 holds with probability $1-\alpha_2$ where $\alpha_2 = n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^{2t}L}\right)$ with distance estimate $\hat{d} = \mathsf{dist}$ defined in (IV.5) with

$$\begin{aligned} d(\theta_u,\theta_v) &= \|\Lambda^{2t+1}Q(e_u - e_v)\|_2^2, \\ \mathsf{bias}(\eta) &= 3B^2|\lambda_r|^{-2t}\sqrt{r\eta}, \\ \Delta &= \Theta(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}), \\ \mathsf{meas}(\eta') &= \frac{(1-\delta)\sqrt{\eta'}}{|\lambda_1|^{2t}L}, \end{aligned} \quad \text{(VI.14)}$$

for any $\eta > 0$, $\delta \in (0,1)$ and $\eta' = \eta - \Delta \in (0, |\lambda_1|^{4t}L^2)$. By substituting the expressions for bias, meas, and α into Lemma 6.1, it follows that for any $\delta' \in (0,1)$,

$$|\hat{F}(u,v,w) - f_r(\theta_u,\theta_v,\theta_w)| \le 3B^2|\lambda_r|^{-2t}\sqrt{r(\eta+\Delta)} + \delta', \quad \text{(VI.15)}$$

with probability at least

$$1 - \exp\left(-\frac{\delta^2(1-\delta)^3 p\left(\sqrt{\eta-\Delta}n\right)^3}{2L^3|\lambda_1|^{6t}}\right) - \exp\left(-\frac{\delta'^2(1-\delta)^4 p\left(\sqrt{\eta-\Delta}n\right)^3}{L^3|\lambda_1|^{6t}}\right) - nO\Big(\exp(-n^{2\psi}(1-o(1))) + n^{-6}\Big) - n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta-\Delta}}{2|\lambda_1|^{2t}L}\right).$$

By selecting

$$\eta = \Delta + \min\left(\Delta, |\lambda_1|^{4t}L^2\right), \quad \text{(VI.16)}$$

it follows by the conditions $\psi > 0$ and $\kappa < \frac{1}{2}$ that

$$\eta + \Delta = \Theta(r\lambda_{\max}^{4t} n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}),$$
$$\eta - \Delta = \Omega(n^{-(\kappa-\psi)}),$$
$$p(\sqrt{\eta-\Delta}n)^3 = \Omega(n^{\frac{5}{4}}),$$
$$n\sqrt{\eta-\Delta} = \Omega(n^{\frac{3}{4}}).$$

By choosing $\delta' = n^{-(\kappa-\psi)/2}$ such that $\delta' = O(\sqrt{\eta})$ and $\delta'^2 p(\sqrt{\eta-\Delta}n)^3 = \Omega(n^{\frac{3}{4}}) = \Omega(n^{2\psi})$ because $\psi < \frac{3}{8}$. Therefore, by substituting into (VI.15), it follows that for any given $(u,v,w) \in [n]^3$, with probability $1 - O(n\exp(-\Theta(n^{2\psi})))\Big) - O(n^{-5})$,

$$\begin{aligned} &|\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w)| \\ &\le |\hat{F}(u,v,w) - f_r(\theta_u,\theta_v,\theta_w)| + |f_r(\theta_u,\theta_v,\theta_w) - f(\theta_u,\theta_v,\theta_w)| \\ &= O\Big(r(\lambda_{\max}/\lambda_r)^{2t} n^{-(\kappa-\psi)/2} + t\varepsilon(1+\varepsilon)^{2t-1} + \sqrt{t\varepsilon(1+\varepsilon)^{2t-1}}\Big), \end{aligned}$$

where the bias between f_r and f is bounded by ε, and dominated by the bound between $\hat{F}$ and f_r. The final result follows from a union bound over $(u,v,w) \in [n]^3$.

The bound on MSE also follows by substituting $\delta = \frac{1}{2}$ and the same choice of η from (VI.16) into Lemma 6.1, and again noting that the bias between F and F_r is dominated by the error between $\hat{F}$ and F_r such that

$$\text{MSE}(\hat{F}) = O\left(r^2(\lambda_{\max}/\lambda_r)^{4t} n^{-(\kappa-\psi)} + t\varepsilon(1+\varepsilon)^{2t-1}\right) + O\left(t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right).$$

This completes the proof of Theorem 5.2.

C. Proof of Corollary 5.3

Proof: The proof follows the same format as the proof of Theorem 5.1. Let us denote the set of anchor vertices as $\mathcal{Y}$ such that $|\mathcal{Y}| = y$, and they are assumed to be chosen uniformly at random amongst all vertices. For a pair of vertices $(a,b) \in \mathcal{Y}^2$, the estimate $\hat{F}(a,b)$ follows the same exact computation as described in Section IV-A. As a result it follows from Theorem 5.1 that with high probability,

$$\max_{(a,b,c)\in\mathcal{Y}^3} |F(a,b,c) - \hat{F}(a,b,c)| = O(n^{-(\kappa-\psi)/2}).$$

Next we need to show the error is not degraded for non-anchor vertices $(u,v,w) \in ([n] \setminus \mathcal{Y})^3$. Let $\zeta : [n] \to \mathcal{Y}$

denote the function that maps from each vertex to the closest anchor vertex as determined by the true distances d,

$$\zeta(u) = \arg\min_{a \in \mathcal{A}} d(\theta_u, \theta_a),$$

and let $\hat{\zeta} : [n] \to \mathcal{Y}$ denote the data-dependent function that maps from each vertex to the closest anchor vertex as determined by the computed distances $\hat{d}$,

$$\hat{\zeta}(u) = \arg\min_{a \in \mathcal{A}} \hat{d}(u, a).$$

The estimate for non-anchor vertices is then taken to be the estimate computed for the corresponding closest anchor vertices,

$$\hat{F}(u, v, w) = \hat{F}(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)),$$

such that

$$\begin{aligned}&|\hat{F}(u, v, w) - F(u, v, w)| \\ &\le |\hat{F}(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w))| \\ &\quad + |F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(u, v, w)|.\end{aligned}$$

By Theorem 5.1, as $(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) \in \mathcal{Y}^3$, the first term is bounded by $O(n^{-(\kappa-\psi)/2})$ with high probability. By property 6.1,

$$\begin{aligned}&|F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(u, v, w)| \\ &\le 3B^2\sqrt{r}|\lambda_r|^{-2t} \\ &\quad \times \sqrt{\max\left(d(\theta_u, \theta_{\hat{\zeta}(u)}), d(\theta_v, \theta_{\hat{\zeta}(v)}), d(\theta_w, \theta_{\hat{\zeta}(w)})\right)}.\end{aligned}$$

The modified algorithm computes distances using Step 4 of the described algorithm between all pairs of anchor vertices, as well as all pairs (u, a) such that $u \in [n]$ and $a \in \mathcal{Y}$. For each computed distance between a pair (u, a), by Lemma 6.2, property 6.2 holds for $\Delta = \Theta(n^{-\kappa+\psi})$ with probability $1 - \alpha_1$ where $\alpha_1 = O\Big(\exp(-n^{2\psi}(1 - o(1)))\Big)$ for $\psi \in (0, \min(\kappa, \frac{3}{8}))$ and $\kappa \in (0, \frac{1}{2})$.

In order to bound $\max_{u \in [n]} d(\theta_u, \theta_{\hat{\zeta}(u)})$, we argue that for every $u \in [n]$, with high probability

$$\begin{aligned}d(\theta_u, \theta_{\hat{\zeta}(u)}) &\overset{(a)}{\le} \hat{d}(u, \hat{\zeta}(u)) + \Delta \\ &\overset{(b)}{\le} \hat{d}(u, \zeta(u)) + \Delta \\ &\overset{(c)}{\le} d(\theta_u, \theta_{\zeta(u)}) + 2\Delta \\ &\overset{(d)}{=} \min_{a \in \mathcal{A}} d(\theta_u, \theta_a) + 2\Delta,\end{aligned}$$

where (a) and (c) hold with high probability for $\Delta = \Theta(n^{-\kappa+\psi})$ as a result of property 6.2, and (b) and (d) follow from the definition of the functions $\hat{\zeta}$ and ζ.

To bound $\min_{a \in \mathcal{A}} d(\theta_u, \theta_a)$, we use (VI.8) from property 6.3 to show that for any $u \in [n]$, $\eta = \Theta(n^{-\kappa+\psi})$, and $y = \Omega((p^2n^3)^{1/4}) = \Omega(n^{\kappa/2})$

$$\begin{aligned}\mathbb{P}\left(\min_{a \in \mathcal{Y}} d(\theta_u, \theta_a) > \eta \mid \theta_u\right) &= \prod_{a \in \mathcal{Y}} \mathbb{P}\left(d(\theta_u, \theta_a) > \eta \mid \theta_u\right) \\ &\le \left(1 - \frac{\sqrt{\eta}}{|\lambda_1|^{2t}L}\right)^y \\ &\le \exp(-\frac{y\sqrt{\eta'}}{|\lambda_1|^{2t}L}) = \exp(-\Theta(n^{\psi/2})).\end{aligned}$$

As a result, the max entrywise error is bounded by $O(n^{-(\kappa-\psi)/2})$ with high probability, which can be used to show the MSE bound of $O(n^{-(\kappa-\psi)})$. □

VII. Proving Distance Estimate Is Close

In this section we argue that the distance estimate as defined in (IV.5) is close to an ideal distance as claimed in the Lemma 6.2.

A. Regular Enough Growth of Breadth-First-Search (BFS) Tree

The distance estimation algorithm of interest constructs a specific BFS tree for each vertex $u \in [n]$ with respect to the bipartite graph between vertices $[n]$ and $\mathcal{V}_A$ where recall that $\mathcal{V}_A = \{(b, c) \in [n/2]^2 \text{ s.t. } b < c\}$. The BFS tree construction is done so that vertices at different levels do not share coordinates, i.e. if vertex $a \in [n]$ is visited in an earlier layer of the BFS tree, then no vertex corresponding to (a, b) for any $b \in [n]$ can be visited subsequently. Similarly, if (a, b) is visited in the BFS tree, then no subsequent vertices including either coordinates a or b can be visited. The restriction is placed across different depths, whereas pairs of vertices (a, b) and (a, c) can be visited in the same depth. Amongst various valid BFS trees, the algorithm chooses one arbitrarily (for example, see Figure 3(c)).

We recall some notations. Consider a valid BFS tree rooted at vertex $u \in [n]$ which respects the constraint that no coordinate is visited more than once. Recall that for any $s \ge 1$, $\mathcal{U}_{u,s} \subseteq \mathcal{V}_A$ denotes the set of vertices at depth $(2s - 1)$ and $\mathcal{S}_{u,s} \subseteq [n]$ denotes the set of vertices at depth $2s$ of the BFS tree, $\mathcal{B}_{u,s} = \cup_{l \in \lceil s/2 \rceil} \mathcal{U}_{u,l} \cup_{h \in \lfloor s/2 \rfloor} \mathcal{S}_{u,h}$, $\mathcal{G}(\mathcal{B}_{u,s})$ denotes all the information corresponding to the subgraph restricted to the first s layers of the BFS tree which includes $\mathcal{B}_{u,s}$, the latent variables $\{\theta_a\}_{a \in \mathcal{B}_{u,s}}$ and the edge weights $\{M_1(a, b, c)\}_{a,(b,c) \in \mathcal{B}_{u,s}}$. The vector $N_{u,s} \in [0, 1]^n$ is such that the a-th coordinate is equal to the product of weights along the path from u to a in the BFS tree for $a \in \mathcal{S}_{u,s}$, and the vector $W_{u,s} \in [0, 1]^{\mathcal{V}_A}$ is such that the (b, c)-th coordinate is equal to the product of weights along the path from u to (b, c) in the BFS tree for $(b, c) \in \mathcal{U}_{u,s}$. The normalized vectors are $\tilde{N}_{u,s} = N_{u,s}/|\mathcal{S}_{u,s}|$ and $\tilde{W}_{u,s} = W_{u,s}/|\mathcal{U}_{u,s}|$ for $u \in [n]$, $s \ge 1$.

In a valid BFS tree rooted at vertex u, $\pi_u(a)$ denotes the parent of $a \in [n]$, and $\pi_u(b, c)$ denotes the parent of $(b, c) \in \mathcal{V}_A$. The neighborhood vectors satisfy recursive relationship,

$$\begin{aligned}N_{u,s}(a) &= M_A(a, \pi_u(a)) W_{u,s}(\pi_u(a)) \mathbb{I}_{(a \in \mathcal{S}_{u,s})} \\ W_{u,s}(b, c) &= M_A(\pi_u(b, c), (b, c)) N_{u,s-1}(\pi_u(b, c)) \mathbb{I}_{((b,c) \in \mathcal{U}_{u,s})}\end{aligned}$$

with $N_{u,0} = e_u$. We state the following result regarding regularity in the growth of the BFS tree.

Lemma 7.1: Let $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$. Let t be as defined in (IV.3). For a given $\delta \in (0, \frac{1}{2})$ and for any $u \in [n]$, with probability $1 - O\Big(n \exp\left(-\Theta(n^{2\kappa})\right)\Big)$, or all $s \in [t - 1]$,

$$|\mathcal{S}_{u,s}| \in \left[(1 - \delta)^{2s} 2^{-3s} n^{2\kappa s}(1 - o(1)), (1 + \delta)^{2s} 2^{-s} n^{2\kappa s}\right], \tag{VII.1}$$

for $s = t$,

$$|\mathcal{S}_{u,t}| \in \left[(1-\delta)^{2t}2^{-3t-1}n^{2\kappa t}(1-o(1)), (1+\delta)^{2t}2^{-t}n^{2\kappa t}\right], \tag{VII.2}$$

and for $s \in [t]$,

$$|\mathcal{U}_{u,s}| \in \Big[(1-\delta)^{2s-1}2^{-3s}n^{\frac{1}{2}+\kappa(2s-1)}(1-o(1)), (1+\delta)^{2s-1}2^{-s}n^{\frac{1}{2}+\kappa(2s-1)}\Big]. \tag{VII.3}$$

The set of single coordinate vertices visited within depth $2t$ is $o(n)$,

$$|\cup_{\ell=0}^{t}\mathcal{S}_{u,\ell}| = o(n). \tag{VII.4}$$

Proof: First observe that if t is as defined in (IV.3) with $\kappa \in (0, \frac{1}{2})$, then

$$t = \left\lceil \frac{\ln(n)}{2\ln(p^2n^3)} \right\rceil = \left\lceil \frac{1}{4\kappa} \right\rceil$$

such that

$$\frac{1}{4\kappa} \le t < \frac{1}{4\kappa} + 1. \tag{VII.5}$$

Note that t is constant with respect to n.

For any $s \in [t]$, we study the growth of $|\mathcal{S}_{u,s}|$ and $|\mathcal{U}_{u,s}|$ conditioned on $\mathcal{B}_{u,2s-1} \cup \mathcal{U}_{u,s}$ and $\mathcal{B}_{u,2(s-1)} \cup \mathcal{S}_{u,s-1}$ respectively. To that end, conditioned on the set $\mathcal{B}_{u,2s-1}$ and the set $\mathcal{U}_{u,s}$, any vertex $i \in [n] \setminus \mathcal{B}_{u,2s-1}$ is in $\mathcal{S}_{u,s}$ independently with probability $(1-(1-p)^{|\mathcal{U}_{u,s}|})$. Thus the number of vertices in $\mathcal{S}_{u,s}$ is distributed as a binomial random variable. By Chernoff's bound,

$$\begin{aligned}&\mathbb{P}\left(|\mathcal{S}_{u,s}| \notin (1\pm\delta)(|[n]\setminus\mathcal{B}_{u,2s-1}|)(1-(1-p)^{|\mathcal{U}_{u,s}|}) \mid \mathcal{B}_{u,2s-1}, \mathcal{U}_{u,s}\right)\\ &\le 2\exp\left(-\frac{1}{3}\delta^2(|[n]\setminus|\mathcal{B}_{u,2s-1}|)(1-(1-p)^{|\mathcal{U}_{u,s}|})\right).\end{aligned} \tag{VII.6}$$

Similarly, conditioned on the sets $\mathcal{B}_{u,2(s-1)}$ and $\mathcal{S}_{u,s-1}$, the set of vertices in $\mathcal{U}_{u,s}$ is equivalent to the number of edges in a graph with vertices $[n/2] \setminus \mathcal{B}_{u,2(s-1)}$ and an edge between (i,j) if there is some $h \in \mathcal{S}_{u,s-1}$ such that $(i,j,h) \in \Omega_1$. This is an Erdos-Renyi graph, as each edge is independent with probability $(1-(1-p)^{|\mathcal{S}_{u,s-1}|})$. By Chernoff's bound,

$$\begin{aligned}&\mathbb{P}\Big(|\mathcal{U}_{u,s}| \notin (1\pm\delta)\tbinom{|[n/2]\setminus\mathcal{B}_{u,2(s-1)}|}{2}\\ &\qquad\times(1-(1-p)^{|\mathcal{S}_{u,s-1}|}) \mid \mathcal{B}_{u,2(s-1)}, \mathcal{S}_{u,s-1}\Big)\\ &\le 2\exp\left(-\frac{1}{3}\delta^2\tbinom{|[n/2]\setminus\mathcal{B}_{u,2(s-1)}|}{2}(1-(1-p)^{|\mathcal{S}_{u,s-1}|})\right).\end{aligned} \tag{VII.7}$$

Let us define the events

$$\mathcal{A}^1_{u,s}(\delta) = \left\{|\mathcal{S}_{u,s}| \in (1\pm\delta)(|[n]\setminus\mathcal{B}_{u,2s-1}|)(1-(1-p)^{|\mathcal{U}_{u,s}|})\right\}, \tag{VII.8}$$

$$\mathcal{A}^2_{u,s}(\delta) = \left\{|\mathcal{U}_{u,s}| \in (1\pm\delta)\tbinom{|[n/2]\setminus\mathcal{B}_{u,2(s-1)}|}{2}(1-(1-p)^{|\mathcal{S}_{u,s-1}|})\right\}. \tag{VII.9}$$

Since $p \in (0,1)$ and hence $1-(1-p)^x \le px$ for all $x \ge 1$, we have that under events $\mathcal{A}^1_{u,s}(\delta) \cap \mathcal{A}^2_{u,s}(\delta)$,

$$|\mathcal{S}_{u,s}| \le (1+\delta)np|\mathcal{U}_{u,s}| \text{ and } |\mathcal{U}_{u,s}| \le (1+\delta)\tbinom{n/2}{2}p|\mathcal{S}_{u,s-1}|,$$

which together implies that conditioned on event $\cap_{h=1}^{s}\big(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta)\big)$, for all $s \in [t]$

$$|\mathcal{S}_{u,s}| \le \left((1+\delta)^2\frac{p^2n^3}{8}\right)^s = (1+\delta)^{2s}2^{-3s}n^{2\kappa s}, \tag{VII.10}$$

and

$$|\mathcal{U}_{u,s}| \le (1+\delta)\frac{pn^2}{8}\left((1+\delta)^2\frac{p^2n^3}{8}\right)^{s-1} \tag{VII.11}$$

$$= (1+\delta)^{2s-1}2^{-3s}n^{\frac{1}{2}+\kappa(2s-1)}. \tag{VII.12}$$

Therefore, for any $s \in [t-1]$ such that $s \le \frac{1}{4\kappa}$ by the definition of t,

$$\begin{aligned}&|\mathcal{B}_{u,2s}|\\ &\le 1+\sum_{\ell=1}^{s}(2|\mathcal{U}_{u,\ell}|+|\mathcal{S}_{u,\ell}|)\\ &\le 1+\sum_{\ell=1}^{s}\left(2(1+\delta)\frac{pn^2}{8}\left((1+\delta)^2\frac{p^2n^3}{8}\right)^{\ell-1}+\left((1+\delta)^2\frac{p^2n^3}{8}\right)^{\ell}\right)\\ &= 1+\left(2(1+\delta)\frac{pn^2}{8}+\left((1+\delta)^2\frac{p^2n^3}{8}\right)\right)\sum_{\ell=1}^{s}\left((1+\delta)^2\frac{p^2n^3}{8}\right)^{\ell-1}\\ &= O\big(pn^2(p^2n^3)^{s-1}\big) = O\big(n^{\kappa(2s-1)+\frac{1}{2}}\big) = O\big(n^{1-\kappa}\big) = o(n).\end{aligned} \tag{VII.13}$$

With a similar argument we can show that

$$\begin{aligned}\sum_{\ell=0}^{t}|\mathcal{S}_{u,\ell}| &\le \sum_{\ell=0}^{t}\left((1+\delta)^2\frac{p^2n^3}{8}\right)^{\ell}\\ &= O((p^2n^3)^t) = O(n^{2\kappa t}) = o(n).\end{aligned} \tag{VII.14}$$

The last step follows from checking that when $\kappa \in [\frac{1}{4}, \frac{1}{2})$, $t = 1$ such that $n^{2\kappa t} = o(n)$, and when $\kappa \in (0, \frac{1}{4})$, from $t \le \frac{1}{4\kappa}+1$, it follows such that $n^{2\kappa t} = O(n^{\frac{1}{2}+2\kappa}) = o(n)$ as $\kappa < \frac{1}{4}$. Recall that we split the coordinates such that $\cup_{\ell=1}^{t}\mathcal{U}_{u,\ell} \subset \mathcal{V}_A$, and the coordinates represented in $(a,b) \in \mathcal{V}_A$ are such that $a \in [n/2]$ and $b \in [n/2]$. Therefore by (VII.14),

$$|[n]\setminus\mathcal{B}_{u,2t-1}| \ge n/2 - \sum_{\ell=0}^{t-1}|\mathcal{S}_{u,\ell}| = \frac{n}{2}(1-o(1)).$$

Using (VII.13), we establish lower bounds on $|\mathcal{S}_{u,s}|$ and $|\mathcal{U}_{u,s}|$ next. Note that, for $p \in (0,1)$, $1-p \le e^{-p}$ and for any $x \in (0,1)$, $e^{-x} \le 1-x+x^2$. It follows that $1-(1-p)^x \ge px(1-px)$. For $s \in [t]$ we can show that

$$\begin{aligned}|\mathcal{U}_{u,s}| &\ge (1-\delta)\frac{(n(1-o(1)))^2(1-o(1))}{8}(1-(1-p)^{|\mathcal{S}_{u,s-1}|})\\ &\ge (1-\delta)\frac{n^2}{8}p|\mathcal{S}_{u,s-1}|(1-p|\mathcal{S}_{u,s-1}|)(1-o(1))\\ &= (1-\delta)\frac{n^2}{8}p|\mathcal{S}_{u,s-1}|(1-o(1)).\end{aligned}$$

For $s \in [t-1]$ we can show that

$$\begin{aligned}|\mathcal{S}_{u,s}| &\ge (1-\delta)n(1-o(1))(1-(1-p)^{|\mathcal{U}_{u,s}|})\\ &\ge (1-\delta)n(1-o(1))p|\mathcal{U}_{u,s}|(1-p|\mathcal{U}_{u,s}|)\\ &\ge (1-\delta)n(1-o(1))p|\mathcal{U}_{u,s}|(1-o(1))\\ &= (1-\delta)pn|\mathcal{U}_{u,s}|(1-o(1)),\end{aligned}$$

and for $s = t$,

$$|\mathcal{S}_{u,t}| \geq (1-\delta)\frac{n}{2}(1-o(1))(1-(1-p)^{|\mathcal{U}_{u,t}|}).$$

Then for $s \in [t]$,

$$\begin{aligned}|\mathcal{U}_{u,s}| &\geq (1-\delta)^2\frac{p^2n^3}{8}|\mathcal{U}_{u,s-1}|(1-o(1))\\ &\geq \left((1-\delta)^2\frac{p^2n^3}{8}\right)^{s-1}|\mathcal{U}_{u,1}|(1-o(1))\\ &\geq \left((1-\delta)^2\frac{p^2n^3}{8}\right)^{s-1}(1-\delta)\frac{pn^2}{8}(1-o(1))\\ &= (1-\delta)^{2s-1}2^{-3s}n^{\frac{1}{2}+\kappa(2s-1)}(1-o(1)); \quad \text{(VII.15)}\end{aligned}$$

for $s \in [t-1]$,

$$\begin{aligned}|\mathcal{S}_{u,s}| &\geq (1-\delta)^2pn\frac{n^2}{8}p|\mathcal{S}_{u,s-1}|(1-o(1)),\\ &\geq \left((1-\delta)^2\frac{p^2n^3}{8}\right)^{s}(1-o(1)),\\ &= (1-\delta)^{2s}2^{-3s}n^{2\kappa s}(1-o(1)); \quad \text{(VII.16)}\end{aligned}$$

and for $s = t$, $|\mathcal{S}_{u,t}| \geq (1-\delta)^{2t}2^{-3t-1}n^{2\kappa t}(1-o(1))$.

To conclude the proof of Lemma 7.1, we need to argue that $\cap_{s=1}^{t}\left(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta)\right)$ holds with high probability. To that end,

$$\begin{aligned}&\mathbb{P}\left(\neg(\cap_{s=1}^{t}(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta)))\right) = \mathbb{P}\left(\cup_{s=1}^{t}\neg(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta))\right)\\ &= \sum_{s=1}^{t}\mathbb{P}\left(\neg(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta))\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right)\\ &\leq \sum_{s=1}^{t}\mathbb{P}\left(\neg(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta))\mid\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right)\\ &\leq \sum_{s=1}^{t}\mathbb{P}\left(\neg\mathcal{A}^1_{u,s}(\delta)\mid\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right)\\ &\quad+\sum_{s=1}^{t}\mathbb{P}\left(\neg\mathcal{A}^2_{u,s}(\delta)\mid\mathcal{A}^1_{u,s}(\delta)\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right).\end{aligned}$$

We bound the each of the two summation terms on the right hand side in the last inequality next. Using (VII.6) and (VII.10), we have

$$\begin{aligned}&\sum_{s=1}^{t}\mathbb{P}\left(\neg\mathcal{A}^1_{u,s}(\delta)\mid\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right)\\ &\leq\sum_{s=1}^{t-1}2\exp\left(-\frac{1}{3}\delta^2(1-\delta)\frac{p^2n^3}{2}\left((1-\delta)^2\frac{p^2n^3}{2}\right)^{s-1}(1-o(1))\right)\\ &\quad+2\exp\left(-\frac{1}{3}\delta^2(1-\delta)\frac{p^2n^3}{4}\left((1-\delta)^2\frac{p^2n^3}{2}\right)^{t-1}(1-o(1))\right)\\ &\leq 4\exp\left(-\frac{1}{3}\delta^2(1-\delta)\frac{p^2n^3}{2}(1-o(1))\right) = O\Big(\exp\big(-\Theta(n^{2\kappa})\big)\Big).\end{aligned}$$

Similarly, using (VII.7) and (VII.12), we have

$$\begin{aligned}&\sum_{s=1}^{t}\mathbb{P}\left(\neg\mathcal{A}^2_{u,s}(\delta)\mid\mathcal{A}^1_{u,s}(\delta)\cap_{h=1}^{s-1}(\mathcal{A}^1_{u,h}(\delta)\cap\mathcal{A}^2_{u,h}(\delta))\right)\\ &\leq\sum_{s=1}^{t+1}2\exp\left(-\frac{1}{3}\delta^2\frac{n^2p}{2}\left((1-\delta)^2\frac{p^2n^3}{2}\right)^{s-1}(1-o(1))\right)\\ &\leq 4\exp\left(-\frac{1}{3}\delta^2\frac{n^2p}{2}(1-o(1))\right)=O\Big(\exp\big(-\Theta(n^{\frac{1}{2}+\kappa})\big)\Big).\end{aligned}$$

Putting it all together, we have that

$$\begin{aligned}&\mathbb{P}\left(\neg(\cap_{s=1}^{t}(\mathcal{A}^1_{u,s}(\delta)\cap\mathcal{A}^2_{u,s}(\delta)))\right)\\ &\leq O\Big(\exp\big(-\Theta(n^{2\kappa})\big)\Big)+O\Big(\exp\big(-\Theta(n^{\frac{1}{2}+\kappa})\big)\Big)\\ &= O\Big(\exp\big(-\Theta(n^{2\kappa})\big)\Big),\end{aligned}$$

since $\kappa \in (0, \frac{1}{2})$. By union bound over all $u \in [n]$, we obtain the desired bound on the probability of error. This concludes the proof of Lemma 7.1. □

B. Concentration of Quadratic Form One

Let $\mathcal{A}^3_{u,t}(\delta)$ denote the event that (VII.1) holds for all $s \in [t-1]$, (VII.2) holds, (VII.3) holds for all $s \in [t]$, and (VII.4) holds. Lemma 7.1 established that this event holds with high probability. Conditioned on the event $\mathcal{A}^3_{u,t}(\delta)$, we prove that a specific quadratic form concentrates around its mean. This will be used as the key property to eventually establish that the distance estimates are a good approximation to the ideal distances.

Lemma 7.2: Let $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$, t as defined in (IV.3), $\delta \in (0, \frac{1}{2})$, and $\psi \in (0, \kappa)$. For any $u \in [n]$, with probability $1 - 2\exp(-n^{2\psi}(1-o(1)))$,

$$|e_k^TQ\tilde{N}_{u,t} - e_k^T\Lambda^{2t}Qe_u| < \frac{16\lambda_k^{2t-2}n^{\psi}}{(1-\delta)n^{\kappa}}.$$

Proof: Recall that conditioning on event $\mathcal{A}^3_{u,t}(\delta)$ simply imposes the restriction that the neighborhood of $u \in [n]$ grows at a specific rate. This event is independent from latent parameters $\{\theta_a\}_{a\in[n]}$, the precise entries in Ω_1 as well as associated values, i.e. M_1.

Conditioned on $\mathcal{A}^3_{u,t}(\delta)$, let $\mathcal{F}_{u,s}$ for $0 \leq s \leq 2t$ denote the sigma-algebra containing information about the latent parameters, edges and the values associated with nodes in the bipartite graph up to distance s from u, i.e. nodes $\mathcal{S}_{u,h'}$ for $h' \leq \lfloor s/2 \rfloor$, $\mathcal{U}_{u,h''}$ for $h'' \leq \lceil s/2 \rceil$, associated latent parameters as well as edges of Ω_1. Specifically, $\mathcal{F}_{u,0}$ contains information about latent parameter θ_u associated with $u \in [n]$; $\mathcal{F}_{u,s}$ contains information about latent parameters $\cup_{h=1}^{\lfloor s/2 \rfloor}\{\theta_a\}_{a\in\mathcal{S}_{u,h}}\cup_{h=1}^{\lceil s/2 \rceil}\{\theta_b,\theta_c\}_{(b,c)\in\mathcal{U}_{u,h}}$ and all the associated edges and observations. This implies that $\mathcal{F}_{u,0} \subset \mathcal{F}_{u,1} \subset \mathcal{F}_{u,2}$, etc.

Recall that Q denotes the $r \times n$ matrix where $Q_{ka} = q_k(\theta_a), k \in [r], a \in [n]$. We modify the notation due to the sample splitting, and we let $\mathcal{Q}$ denote the $r \times \binom{n/2}{2}$ matrix where $\mathcal{Q}_{kb} = q_k(\theta_{b_1})q_k(\theta_{b_2})$ for some $b \in \mathcal{V}_A$ that represents the pair of coordinates (b_1, b_2) for $b_1 < b_2 \in [n/2]$.

We shall consider a specific martingale sequence with respect to the filtration $\mathcal{F}_{u,s}$ that will help establish the desired concentration of $e_k^TQ\tilde{N}_{u,t} - e_k^T\Lambda^{2t}Qe_u$. For $1 \leq s \leq 2t$, define

$$\begin{aligned}Y_{u,s} &= \begin{cases}e_k^T\Lambda^{2t-s}Q\tilde{N}_{u,s/2} & \text{if } s \text{ even}\\ e_k^T\Lambda^{2t-s}\mathcal{Q}\tilde{W}_{u,(s+1)/2} & \text{if } s \text{ odd}\end{cases}\\ D_{u,s} &= Y_{u,s} - Y_{u,s-1},\end{aligned}$$

$$Y_{u,2t} - Y_{u,0} = e_k^TQ\tilde{N}_{u,t} - e_k^T\Lambda^{2t}Q\tilde{N}_{u,0} = \sum_{s=1}^{2t}D_{u,s}.$$

Note that $\tilde{N}_{u,0} = e_u$, and $Y_{u,s}$ is measurable with respect to $\mathcal{F}_{u,s}$ because $e_k^T \Lambda^{2t-s} Q \tilde{N}_{u,s/2}$ and $e_k^T \Lambda^{2t-s} Q \tilde{N}_{u,(s+1)/2}$ only depend on observations in the BFS tree within depth s.

By Lemmas 7.4 and 7.5, it follows that $Y_{u,s}$ is martingale with respect to $\mathcal{F}_{u,s}$ for $1 \le s \le t$, i.e.

$$\mathbb{E}[D_{u,s} \mid \mathcal{F}_{u,s-1}] = 0. \tag{VII.17}$$

Furthermore, for properly chosen ν_s as specified in Lemmas 7.4 and 7.5,

$$\mathbb{E}[e^{\lambda D_s} \mid \mathcal{F}_{s-1}, \mathcal{A}^3_{u,t}(\delta)] \le e^{\lambda^2 \nu_s^2/2}$$

almost surely for any $\lambda \in \mathbb{R}$.

We can then apply Proposition 7.3 with any arbitrarily small α_* such that for any $x \ge 0$,

$$\mathbb{P}\left(|e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^{2t} Q e_u| \ge x \mid \mathcal{A}^3_{u,t}(\delta)\right) \le 2\exp\left(-\frac{x^2}{2\sum_{s=1}^{2t} \nu_s^2}\right)$$

where for n large enough,

$$\begin{aligned}\sum_{s=1}^{2t} \nu_s^2 &= \sum_{s=1}^{t} \frac{(1+4\pi)\lambda_k^{4t-4s} 2^{3s-1}(1+o(1))}{(1-\delta)^{2s} n^{2\kappa s}} \\ &\quad + \sum_{s=1}^{t} \frac{(1+16\pi)72\lambda_k^{4t-4s+2} B^4 2^{3s-1}(1+o(1))}{(1-\delta)^{2s-1} n^{\min\{1,\frac{1}{2}+\kappa(2s-1)\}}} \\ &\le \frac{8(1+4\pi)\lambda_k^{4t-4}(1+o(1))}{(1-\delta)^2 n^{2\kappa}},\end{aligned}$$

and $1 + 4\pi \le 16$.

For $\psi \in (0, \kappa)$, we choose $x = \frac{16\lambda_k^{2t-2} n^{\psi}}{(1-\delta)n^{\kappa}} = o(1)$, such that with probability $1 - 2\exp(-n^{2\psi}(1-o(1)))$,

$$|e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^{2t} Q e_u| < x.$$

□

We recall the following concentration inequality for Martingale difference sequence, cf. [46, Theorem 2.19]:

Proposition 7.3: Let $\{D_k, \mathcal{F}_k\}_{k \ge 1}$ be a martingale difference sequence such that $\mathbb{E}[e^{\lambda D_k} | \mathcal{F}_{k-1}] \le e^{\lambda^2 \nu_k^2/2}$ almost surely for all $\lambda \in \mathbb{R}$. Then for all $x > 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \ge x\right) \le 2\exp\left(-\frac{x^2}{2\sum_{k=1}^{n} \nu_k^2}\right). \tag{VII.18}$$

Lemma 7.4: For any $s \in [t]$,

$$\mathbb{E}\left[D_{u,2s} \mid \mathcal{F}_{u,2s-1}, \mathcal{A}^3_{u,t}(\delta)\right] = 0.$$

Let $\nu = \sqrt{\frac{Q(1+4\pi)}{2}}$, and

$$Q = \frac{B^2 \lambda_k^{4t-4s} 2^{3s}(1+o(1))}{(1-\delta)^{2s} n^{2\kappa s}}.$$

For any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda D_{u,2s}} \mid \mathcal{F}_{u,2s-1}, \mathcal{A}^3_{u,t}(\delta)\right] \le \exp\left(\frac{\lambda^2 \nu^2}{2}\right).$$

Proof: Recall $\mathcal{F}_{2s-1}$ contains all information in the depth $2s-1$ neighborhood of vertex u. In particular this includes the vertex set

$$\mathcal{B}_{u,2s-1} = \cup_{l=1}^{s} \mathcal{U}_{u,l} \cup_{h=1}^{s-1} \mathcal{S}_{u,h},$$

the vertex latent variables $\{\theta_i\}_{i \in \mathcal{B}_{u,2s-1}}$ and the edges and corresponding weights. Let us additionally condition on the set $\mathcal{S}_{u,s}$. As $2s$ is even,

$$\begin{aligned}D_{u,2s} &= Y_{u,2s} - Y_{u,2s-1} \\ &= \lambda_k^{2t-2s}\left(e_k^T Q \tilde{N}_{u,s} - \lambda_k e_k^T \mathcal{Q} \tilde{W}_{u,s}\right) \\ &= \lambda_k^{2t-2s}\left(\frac{1}{|\mathcal{S}_{u,s}|}\sum_{i \in [n]} N_{u,s}(i) q_k(\theta_i) - \lambda_k e_k^T \mathcal{Q} \tilde{W}_{u,s}\right) \\ &= \lambda_k^{2t-2s}\left(\frac{1}{|\mathcal{S}_{u,s}|}\sum_{i \in \mathcal{S}_{u,s}} \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} \right. \\ &\qquad \left. \times W_{u,s}(a)\mathbb{I}_{(a=\pi(i))} M_1(a_1, a_2, i) q_k(\theta_i) - \lambda_k e_k^T \mathcal{Q} \tilde{W}_{u,s}\right).\end{aligned}$$

Let us define

$$\begin{aligned}X_i &= \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} W_{u,s}(a)\mathbb{I}_{(a=\pi(i))} M_1(a_1, a_2, i) q_k(\theta_i) \\ &= \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} W_{u,s}(a)\mathbb{I}_{(a=\pi(i))}(f(\theta_{a_1}, \theta_{a_2}, \theta_i) + \epsilon_{a_1 a_2 i}) q_k(\theta_i).\end{aligned}$$

The randomness in X_i only depends on $\theta_i, \epsilon_{a_1 a_2 i}, \mathbb{I}_{(a=\pi(i))}$. Note that we already conditioned on $\theta_{a_1}, \theta_{a_2}$ for $a \in \mathcal{U}_{u,s} \subset \mathcal{B}_{2s-1}$. X_i is independent from X_j because the vertices and edges are disjoint, and $\pi(i)$ is independent from $\pi(j)$ as different vertices are allowed to have the same (or different) parents. First we compute the mean of X_i (conditioned on $i \in \mathcal{S}_{u,s}$). For any vertex $i \in \mathcal{S}_{u,s}$, it must have exactly one parent in $\mathcal{U}_{u,s}$ due to the BFS tree constraints. The parent is equally likely to be any vertex in $\mathcal{U}_{u,s}$ due to the symmetry in the randomly sampled observations. Because the additive noise terms are mean zero, the eigenfunctions are orthonormal, and $\pi(i)$ is equally likely to be any $a \in \mathcal{U}_{u,s}$, it follows that $\mathbb{E}[X_i | i \in \mathcal{S}_{u,s}] = e_k^T \Lambda \mathcal{Q} \tilde{W}_{u,s}$ as shown in (VII.19) shown at the bottom of the next page.

Furthermore, $|X_i| \le B$ almost surely as we assumed $|q_k(\theta)| \le B$. By Hoeffding's inequality, it follows that

$$\mathbb{P}\left(|D_{u,2s}| \ge z \mid \mathcal{F}_{u,2s-1}, \mathcal{S}_{u,s}\right) \le 2\exp\left(-\frac{2|\mathcal{S}_{u,s}|z^2}{\lambda_k^{4t-4s} B^2}\right).$$

If we condition on the event $\mathcal{A}^3_{u,t}(\delta)$,

$$|\mathcal{S}_{u,s}| \ge (1-\delta)^{2s} 2^{-3s-1} n^{2\kappa s}(1-o(1))$$

for $s \in [t]$. Therefore,

$$\mathbb{P}\left(|D_{u,2s}| \ge z \mid \mathcal{F}_{u,2s-1}, \mathcal{A}^3_{u,t}(\delta)\right) \le 2\exp\left(-\frac{2(1-\delta)^{2s} 2^{-3s-1} n^{2\kappa s}(1-o(1))z^2}{\lambda_k^{4t-4s} B^2}\right).$$

We finish the proof by using Lemma A.2 with $c = 2$ and $Q = \frac{B^2 \lambda_k^{4t-4s} 2^{3s}(1+o(1))}{(1-\delta)^{2s} n^{2\kappa s}}$. □

Lemma 7.5: For any $s \in [t]$,

$$\mathbb{E}\left[D_{u,2s-1} \mid \mathcal{F}_{u,2s-1}, \mathcal{A}^3_{u,t}(\delta)\right] = 0.$$

Let $\nu = \sqrt{\frac{Q(1+16\pi)}{2}}$, and

$$Q = \frac{72\lambda_k^{4t-4s+2} B^4 2^{3s}(1+o(1))}{(1-\delta)^{2s-1} n^{\min\{1,\frac{1}{2}+\kappa(2s-1)\}}}.$$

For any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda D_{u,2s-1}} \mid \mathcal{F}_{u,2s-1}, \mathcal{A}^3_{u,t}(\delta)\right] \le \exp\left(\frac{\lambda^2 \nu^2}{2}\right).$$

Proof: As $2s-1$ is odd,

$$D_{u,2s-1} = Y_{u,2s-1} - Y_{u,2(s-1)} = \lambda_k^{2t-2s+1}\left(e_k^T \mathcal{Q}\tilde{W}_{u,s} - \lambda_k e_k^T Q \tilde{N}_{u,s-1}\right).$$

Recall $\mathcal{F}_{2s-2}$ contains all information in the depth $2s-2$ neighborhood of vertex u. In particular this includes the vertex set

$$\mathcal{B}_{u,2s-2} = \cup_{l\in[s-1]}\mathcal{U}_{u,l} \cup_{h\in[s-1]} \mathcal{S}_{u,h},$$

the vertex latent variables $\{\theta_i\}_{i\in\mathcal{B}_{u,2(s-1)}}$ and the edges and corresponding weights $\{M_1(i,a)\}_{i,a\in\mathcal{B}_{u,2(s-1)}}$. Consider

$$\begin{aligned} &e_k^T \mathcal{Q}\tilde{W}_{u,s} \\ &= \frac{1}{|\mathcal{U}_{u,s}|}\sum_{i=(i_1,i_2)\in\mathcal{U}_{u,s}} W_{u,s}(i) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \\ &= \frac{1}{|\mathcal{U}_{u,s}|}\sum_{i=(i_1,i_2)\in\mathcal{U}_{u,s}}\sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\mathbb{I}_{(v=\pi(i))} \\ &\quad\times M_1(v,i) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \\ &= \frac{1}{|\mathcal{U}_{u,s}|}\sum_{i=(i_1,i_2)\in\mathcal{U}_{u,s}} X_i, \end{aligned}$$

where we define for $i=(i_1,i_2)\in\mathcal{U}_{u,s}$,

$$\begin{aligned} X_i &= \sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\mathbb{I}_{(v=\pi(i))} M_1(v,i) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \\ &= \sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\mathbb{I}_{(v=\pi(i))} \\ &\quad\times\left(\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) + \epsilon_{v i_1 i_2}\right) q_k(\theta_{i_1}) q_k(\theta_{i_2}). \end{aligned}$$

Conditioned on $\mathcal{F}_{u,2(s-1)}$ the randomness in X_i only depends on $\theta_{i_1},\theta_{i_2}$, $\epsilon_{\pi(i)i_1i_2}$, and $\mathbb{I}_{(v=\pi(i))}$. Conditioned on $\mathcal{U}_{u,s}$ and $\{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{U}_{u,s}}$, the random variables X_i are independent as $\epsilon_{\pi(i)i_1i_2}$ and $\mathbb{I}_{(v=\pi(i))}$ are independent. The parent of $i=(i_1,i_2)\in\mathcal{U}_{u,s}$ is equally likely to be any vertex in $\mathcal{S}_{u,s-1}$, and the parent of $i=(i_1,i_2)\in\mathcal{U}_{u,s}$ is independent from the parent of $j=(j_1,j_2)\in\mathcal{U}_{u,s}$ with $j\neq i$ as different vertices are allowed to have the same (or different) parent. First we compute the mean of X_i conditioned on $i\in\mathcal{U}_{u,s}$ and $\theta_{i_1},\theta_{i_2}$. Because the additive noise terms are mean zero and the parent of i is equally likely to be any $v\in\mathcal{S}_{u,s-1}$,

$$\begin{aligned} &\mathbb{E}[X_i \mid \theta_{i_1},\theta_{i_2}, i\in\mathcal{U}_{u,s}] \\ &= \mathbb{E}\Big[\sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\mathbb{I}_{(v=\pi(i))}\left(\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2})\right) \\ &\qquad\times q_k(\theta_{i_1}) q_k(\theta_{i_2}) \;\Big|\; \theta_{i_1},\theta_{i_2}\Big] \\ &= \frac{1}{|\mathcal{S}_{u,s-1}|}\sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\left(\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2})\right) \\ &\quad\times q_k(\theta_{i_1}) q_k(\theta_{i_2}). \end{aligned}$$

Furthermore, $|X_i| \le B^2$ almost surely as we assumed $|q_k(\theta)|\le B$. By Hoeffding's inequality, it follows that

$$\begin{aligned} &\mathbb{P}\left(|e_k^T\mathcal{Q}\tilde{W}_{u,s} - \mathbb{E}[e_k^T\mathcal{Q}\tilde{W}_{u,s} \mid \{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{U}_{u,s}},\mathcal{U}_{u,s}]| \right. \\ &\qquad \left. > z \;\Big|\; \mathcal{F}_{u,2(s-1)},\mathcal{U}_{u,s},\{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{U}_{u,s}}\right) \\ &\le 2\exp\left(-\frac{|\mathcal{U}_{u,s}|z^2}{B^4}\right). \end{aligned} \tag{VII.20}$$

Next we consider concentration with respect to the random subset $\mathcal{U}_{u,s}$ out of the $\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}$ possible vertices. In particular we would like to argue that with high probability,

$$\begin{aligned} &\frac{1}{|\mathcal{U}_{u,s}|}\sum_{i\in\mathcal{U}_{u,s}}\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \\ &\quad\approx \frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|} \\ &\qquad\times\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}}\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}). \end{aligned}$$

Next, we formalize it. To that end, conditioned on the size $|\mathcal{U}_{u,s}|$, the set $\mathcal{U}_{u,s}$ is a uniform random sample of the possible set of vertices $\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}$. The above expression on the left is thus the mean of a random sample $\mathcal{U}_{u,s}$ without replacement from $\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}$. Due to negative dependence, it concentrates around its means no slower than assuming that they were a sample of the same size from the same population with replacement, cf. [47, Theorem 4]. Therefore, using

$$\begin{aligned} &\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \\ &\quad = |f(\theta_v,\theta_{i_1},\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2})| \le B^2, \end{aligned}$$

$$\begin{aligned} &\mathbb{E}[X_i \mid i\in\mathcal{S}_{u,s}] \\ &= \mathbb{E}\left[\sum_{a=(a_1,a_2)\in\mathcal{U}_{u,s}} W_{u,s}(a)\mathbb{I}_{(a=\pi(i))}(f(\theta_{a_1},\theta_{a_2},\theta_i)+\epsilon_{a_1a_2i}) q_k(\theta_i)\right] \\ &= \sum_{a=(a_1,a_2)\in\mathcal{U}_{u,s}} \frac{1}{|\mathcal{U}_{u,s}|}\mathbb{E}\left[W_{u,s}(a)\sum_h \lambda_h q_h(\theta_{a_1}) q_h(\theta_{a_2}) q_h(\theta_i) q_k(\theta_i)\right] \\ &= \sum_{a=(a_1,a_2)\in\mathcal{U}_{u,s}} \tilde{W}_{u,s}(a)\lambda_k q_k(\theta_{a_1}) q_k(\theta_{a_2}) \\ &= e_k^T\Lambda\mathcal{Q}\tilde{W}_{u,s}. \end{aligned} \tag{VII.19}$$

we can apply Hoeffding's inequality to argue that

$$\begin{aligned}&\mathbb{P}\Big(\Big|\frac{1}{|\mathcal{U}_{u,s}|}\sum_{i\in\mathcal{U}_{u,s}} q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\\&\quad-\frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|}\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}} q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\Big|\\&\quad\geq z\mid\{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{V}_A},|\mathcal{U}_{u,s}|\Big)\\&\leq 2\exp\Big(-\frac{|\mathcal{U}_{u,s}|z^2}{B^4}\Big).\end{aligned}\tag{VII.21}$$

Finally, we need to account for the randomness in $\{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{V}_A}$, arguing that with high probability

$$\begin{aligned}&\frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|}\\&\quad\times\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}}\sum_l \lambda_l q_l(\theta_v)q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\\&\approx\lambda_k q_k(\theta_v).\end{aligned}$$

To formalize this, we start by recalling that

$$\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}=\{(i_1,i_2)\text{ s.t. }i_1<i_2,\{i_1,i_2\}\subset[n/2]\setminus\mathcal{B}_{u,2(s-1)}\}.$$

Let $n_{u,s}=|[n/2]\setminus\mathcal{B}_{u,2(s-1)}|$, then $|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|=\binom{|n_{u,s}|}{2}$. Then the above summation can be written as a pairwise U-statistic,

$$U=\frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|}\sum_{(i_1,i_2)\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}} g(\theta_{i_1},\theta_{i_2})$$

where g is a symmetric function and each term $g(\theta_{i_1},\theta_{i_2})$ is bounded in absolute value by B^2. Furthermore,

$$\mathbb{E}[\sum_l \lambda_l q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})]=\lambda_k q_k(\theta_v)$$

by the orthogonality model assumption. Therefore, by Lemma A.3

$$\begin{aligned}&\mathbb{P}\Big(\Big|\frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|}\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}}\sum_l \lambda_l q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\\&\qquad-\lambda_k q_k(\theta_v)\Big|\geq z\Big)\\&\leq 2\exp\Big(-\frac{n_{u,s}z^2}{8B^4}\Big).\end{aligned}\tag{VII.22}$$

By putting together all calculations, it also follows that

$$\mathbb{E}[e_k^T\mathcal{Q}\tilde{W}_{u,s}]=e_k^T\Lambda Q\tilde{N}_{u,s-1},$$

and for $z_1,z_2,z_3>0$, with probability at least

$$1-2\exp\Big(-\frac{|\mathcal{U}_{u,s}|z_1^2}{B^4}\Big)-2\exp\Big(-\frac{|\mathcal{U}_{u,s}|z_2^2}{B^4}\Big)-2\exp\Big(-\frac{n_{u,s}z_3^2}{8B^4}\Big)$$

it holds that $|e_k^T\mathcal{Q}\tilde{W}_{u,s}-e_k^T\Lambda Q\tilde{N}_{u,s-1}|\leq z_1+z_2+z_3$ as shown in (VII.23), shown at the bottom of the next page, using the fact that $\|N_{u,s-1}\|_\infty\leq 1$. Conditioned on $\mathcal{A}^3_{u,t}(\delta)$, $n_{u,s}=n/2(1-o(1))$, and

$$\begin{aligned}|\mathcal{U}_{u,s}|\in\Big[&(1-\delta)^{2s-1}2^{-3s}n^{\frac{1}{2}+\kappa(2s-1)}(1-o(1)),\\&(1+\delta)^{2s-1}2^{-s}n^{\frac{1}{2}+\kappa(2s-1)}\Big].\end{aligned}$$

As a result, for $z_1=z_2=z_3$, the expression in (VII.20) and (VII.21) asymptotically dominate the expression in (VII.22). It follows that, with appropriate choice of $z_1=z_2=z_3$ in the above,

$$\begin{aligned}&\mathbb{P}\left(|D_{u,2s-1}|\geq z\mid\mathcal{F}_{2s-2},\mathcal{A}^3_{u,t}(\delta)\right)\\&\leq 6\exp\left(-\frac{(1-\delta)^{2s-1}2^{-3s}n^{\min\{1,\frac{1}{2}+\kappa(2s-1)\}}(1-o(1))z^2}{72\lambda_k^{4t-4s+2}B^4}\right).\end{aligned}$$

We finish the proof by using Lemma A.2 with $c=4$ and $Q=\frac{72\lambda_k^{4t-4s+2}B^4 2^{3s}(1+o(1))}{(1-\delta)^{2s-1}n^{\min\{1,\frac{1}{2}+\kappa(2s-1)\}}}$. □

C. Concentration of Quadratic Form Two

Lemma 7.2 suggests the following high probability events: for any $u\in[n],k\in[r]$, t as defined in (IV.3), i.e. $t=\left\lceil\frac{1}{4\kappa}\right\rceil$, $\delta\in(0,1)$, and

$$x=\frac{16\lambda_{\max}^{2t-2}n^{\psi}}{(1-\delta)n^{\kappa}}.$$

define

$$\mathcal{A}^4_{u,k,t}(x,\delta)=\left\{|e_k^TQ\tilde{N}_{u,t}-e_k^T\Lambda^{2t}Qe_u|<x\right\}\cap\mathcal{A}^3_{u,t}(\delta).$$

Now, we state a useful concentration that builds on the above condition holding. It will be useful step towards establishing Lemma 6.2.

Lemma 7.6: Let $p=n^{-3/2+\kappa}$ for $\kappa\in(0,\frac{1}{2})$, t as defined in (IV.3), and $\delta\in(0,\frac{1}{2})$. For any $u,v\in[n]$, conditioned on $\cap_{k=1}^r\big(\mathcal{A}^4_{u,k,t}(x,\delta)\cap\mathcal{A}^4_{v,k,t}(x,\delta)\big)$,

$$\begin{aligned}&\left|\tilde{N}_{u,t}^TQ^T\Lambda^2Q\tilde{N}_{v,t}-e_u^TQ^T\Lambda^{2(2t+1)}Qe_v\right|\\&\qquad\leq x^2\left(\sum_{k=1}^r\lambda_k^2\right)+xB\left(\sum_{k=1}^r 2\lambda_k^{2(t+1)}\right).\end{aligned}$$

Proof: Proof of Lemma 7.6. Assuming event $\cap_{k=1}^r\big(\mathcal{A}^4_{u,k,t}(x,\delta)\cap\mathcal{A}^4_{v,k,t}(x,\delta)\big)$ holds,

$$\begin{aligned}&|\tilde{N}_{u,t}^TQ^T\Lambda^2Q\tilde{N}_{v,t}-e_u^TQ^T\Lambda^{2(2t+1)}Qe_v|\\&\leq|(\tilde{N}_{u,t}^TQ^T-e_u^TQ^T\Lambda^{2t})(\Lambda^2Q\tilde{N}_{v,t}-\Lambda^{2(t+1)}Qe_v)|\\&\quad+|(\tilde{N}_{u,t}^TQ^T-e_u^TQ^T\Lambda^{2t})\Lambda^{2(t+1)}Qe_v|\\&\quad+|e_u^TQ^T\Lambda^{2(t+1)}(Q\tilde{N}_{v,t}-\Lambda^{2t}Qe_v)|\\&\leq\Big|\sum_{k=1}^r(e_k^TQ\tilde{N}_{u,t}-e_k^T\Lambda^{2t}Qe_u)(e_k^T\Lambda^2Q\tilde{N}_{v,t}-e_k^T\Lambda^{2(t+1)}Qe_v)\Big|\\&\quad+\Big|\sum_{k=1}^r(e_k^TQ\tilde{N}_{u,t}-e_k^T\Lambda^{2t}Qe_u)e_k^T\Lambda^{2(t+1)}Qe_v\Big|\\&\quad+\Big|\sum_{k=1}^r(e_k^T\Lambda^{2(t+1)}Qe_u)(e_k^TQ\tilde{N}_{v,t}-e_k^T\Lambda^{2t}Qe_v)\Big|.\end{aligned}\tag{VII.24}$$

In above, we have simply used the fact that for two vectors $a,b\in\mathbb{R}^r$, $a^Tb=\sum_k a_kb_k=\sum_k(e_k^Ta)(e_k^Tb)$. Now, consider the first term on the right hand side of the last inequality. If $\cap_{k=1}^r\mathcal{A}^4_{u,k,t}(x,\delta)$ holds, then $|(e_k^TQ\tilde{N}_{u,t}-e_k^T\Lambda^{2t}Qe_u)|\leq x$. And if $\cap_{k=1}^r\mathcal{A}^4_{v,k,t}(x,\delta)$ holds, then $|(e_k^T\Lambda^2Q\tilde{N}_{v,t}-e_k^T\Lambda^{2(t+1)}Qe_v)|\leq\lambda_k^2x$. Similar application to other terms and the fact that $|e_k^TQe_u|,|e_k^TQe_v|\leq\|q_k(\cdot)\|_\infty\leq B$, we conclude that

$$\begin{aligned}&|\tilde{N}_{u,t}^TQ^T\Lambda^2Q\tilde{N}_{v,t}-e_u^TQ^T\Lambda^{2(2t+1)}Qe_v|\\&\leq x^2\left(\sum_{k=1}^r\lambda_k^2\right)+xB\left(\sum_{k=1}^r 2\lambda_k^{2(t+1)}\right).\end{aligned}\tag{VII.25}$$

□

D. Concentration of Quadratic Form Three

We establish a final concentration that will lead us to the proof of good distance function property. For any $u \in [n]$, define event

$$A'_{u,v,t}(x,\delta) = \cap_{k=1}^{r}\big(\mathcal{A}^4_{u,k,t}(x,\delta) \cap \mathcal{A}^4_{v,k,t}(x,\delta)\big). \quad \text{(VII.26)}$$

Lemma 7.7: Let $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$, t as defined in (IV.3), $\delta \in (0, \frac{1}{2})$, and

$$x = \frac{16\lambda_{\max}^{2t-2} n^{\psi}}{(1-\delta)n^{\kappa}}.$$

Let $S \equiv S_{u,v,t} = [n]\backslash(\mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t} \cup [n/2])$. Then, under event $A'_{u,v,t}(x,\delta)$,

$$\left|\frac{1}{\binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{\alpha<\beta\in S\times S} T(\alpha,\beta) - \tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t}\right| = O\left(\frac{n^{\psi}}{(|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}}\right) + O\left(\frac{n^{\psi}}{|S|^{1/2}}\right)$$

with probability at least $1-4\exp(-n^{2\psi}(1-o(1)))-O(n^{-6})$ with $\psi \in (0,\kappa)$.

Proof: First, note that $A'_{u,v,t}(x,\delta)$ includes events $\mathcal{A}^3_{u,t}(\delta)$ and $\mathcal{A}^3_{v,t}(\delta)$. This implies that $|S| = \frac{n}{2} - o(n) = \frac{n(1-o(1))}{2}$. Furthermore, it implies that $|\mathcal{S}_{u,t}|$ and $|\mathcal{S}_{v,t}|$ are both greater than or equal to $(1-\delta)^{2t}2^{-3t-1}n^{2\kappa t}(1-o(1))$. As a result,

$$|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}| \geq \frac{n^2}{8}p^2\left(\frac{(1-\delta)^2}{8}n^{2\kappa}\right)^{2t}(1-o(1)) \quad \text{(VII.27)}$$

$$\geq n^{2+4\kappa t}n^{2(-\frac{3}{2}+\kappa)}\frac{1}{8}\left(\frac{(1-\delta)^2}{8}\right)^{2t}(1-o(1)) \quad \text{(VII.28)}$$

$$= \Theta(n^{-1+2\kappa(2t+1)}) \quad \text{(VII.29)}$$

$$= \Omega(n^{2\kappa}). \quad \text{(VII.30)}$$

The asymptotic relationships follow from the choice of $t \geq \frac{1}{4\kappa}$, and the fact that δ and t are both constants.

Recall that $M_B(a,(\alpha,\beta)) = \mathbb{I}_{((a,\alpha,\beta)\in\Omega_1)}(F(a,\alpha,\beta) + \epsilon_{a\alpha\beta})$ for

$$F(a,\alpha,\beta) = \sum_{k=1}^{r}\lambda_k q_k(\theta_a)q_k(\theta_\alpha)q_k(\theta_\beta).$$

There are 3 sources of randomness: the sampling of entries in Ω_1, the observation noise terms $\epsilon_{a\alpha\beta}$, and the latent variables $\theta_a, \theta_\alpha, \theta_\beta$. Since we enforce that α and β are in the complement of $\mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}$, the sampling, observations, and latent variables involved in M_B are independent from $N_{u,t}$ and $N_{v,t}$.

Let us define the quantity

$$\begin{aligned}\tilde{T}(\alpha,\beta) &= \min(\max(T(\alpha,\beta), -\phi^2), \phi^2)\\ &= \text{sign}(T(\alpha,\beta))\min(|T(\alpha,\beta)|, \phi^2)\end{aligned}$$

for $\phi = \lceil 16/(1-2\kappa)\rceil$ where recall that

$$T(\alpha,\beta) = \sum_{a\neq b\in[n]} N_{u,t}(a)N_{v,t}(b)M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta)).$$

Trivially, due to this thresholding, $|\tilde{T}(\alpha,\beta)| \leq \phi^2$ such that $|\tilde{T}(\alpha,\beta) - \mathbb{E}[\tilde{T}(\alpha,\beta)]| \leq 2\phi^2$.

To begin with, $N_{u,t}(a) = 0$ if $a \notin \mathcal{S}_{u,t} \subset \mathcal{B}_{u,2t}$ and $N_{v,t}(b) = 0$ if $b \notin \mathcal{S}_{v,t} \subset \mathcal{B}_{v,2t}$. Further, conditioned on event $A'_{u,v,t}(x,\delta)$, all the information associated with $\mathcal{B}_{u,2t}$ and $\mathcal{B}_{v,2t}$ is revealed; however, information about $[n]\backslash(\mathcal{B}_{u,2t}\cup\mathcal{B}_{v,2t})$ is not. Let $\mathcal{F}(u,v,t,x,\delta)$ denote all the information revealed such that event $A'_{u,v,t}(x,\delta)$ holds.

Let's prove concentration in two steps. In step one, we condition on $\mathcal{F}(u,v,t,x,\delta)$ and the latent variables $\{\theta_i\}_{i\in[n]}$. The sampling process (edges in Ω_1) and the observation noise are independent for distinct pairs (α,β) and (α',β'). As a result, $T(\alpha,\beta)$ and $T(\alpha',\beta')$ are conditionally independent as long as $\{\alpha,\beta\}\cap\{\alpha',\beta'\} \neq 2$, i.e. they are not the exact same pair. The correlations across $T(\alpha,\beta)$ and $T(\alpha',\beta')$ are due only to the latent variables if $\alpha,\beta,\alpha',\beta'$ share any values. We will bound the variance of $T(\alpha,\beta)$ in Lemma 7.8, and by combining it with the conditional independence property across $T(\alpha,\beta)$,

$$\begin{aligned}
&|e_k^T\mathcal{Q}\tilde{W}_{u,s} - e_k^T\Lambda Q\tilde{N}_{u,s-1}|\\
&\leq |e_k^T\mathcal{Q}\tilde{W}_{u,s} - \mathbb{E}[e_k^T\mathcal{Q}\tilde{W}_{u,s} \mid \{\theta_{i_1},\theta_{i_2}\}_{i\in\mathcal{U}_{u,s}},\mathcal{U}_{u,s}]|\\
&+ \Big|\frac{1}{|\mathcal{S}_{u,s-1}|}\sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\Big(\frac{1}{|\mathcal{U}_{u,s}|}\sum_{i\in\mathcal{U}_{u,s}}\sum_l \lambda_l q_l(\theta_v)q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\\
&\qquad - \frac{1}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|}\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}}\sum_l \lambda_l q_l(\theta_v)q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})\Big)\Big|\\
&+ \Big|\frac{1}{|\mathcal{S}_{u,s-1}|}\sum_{v\in\mathcal{S}_{u,s-1}} N_{u,s-1}(v)\Big(\frac{\sum_{i\in\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}}\sum_l \lambda_l q_l(\theta_v)q_l(\theta_{i_1})q_l(\theta_{i_2})q_k(\theta_{i_1})q_k(\theta_{i_2})}{|\mathcal{V}_A\setminus\mathcal{B}_{u,2(s-1)}|} - \lambda_k q_k(\theta_v)\Big)\Big|\\
&\leq z_1 + \frac{1}{|\mathcal{S}_{u,s-1}|}\sum_{v\in\mathcal{S}_{u,s-1}}|N_{u,s-1}(v)|z_2 + \frac{1}{|\mathcal{S}_{u,s-1}|}\sum_{v\in\mathcal{S}_{u,s-1}}|N_{u,s-1}(v)|z_3\\
&\leq z_1 + z_2 + z_3, \qquad \text{(VII.23)}
\end{aligned}$$

it follows that (using notation $\mathcal{F} = \mathcal{F}(u, v, t, x, \delta)$)

$$\begin{aligned}
&\text{Var}\left[\sum_{\alpha<\beta\in S\times S} T(\alpha,\beta) \;\middle|\; \mathcal{F}, \{\theta_i\}_{i\in[n]}\right] \\
&\quad= \sum_{\alpha<\beta\in S\times S} \text{Var}\left[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}\right] \\
&\quad\le 2\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|(1+o(1)).
\end{aligned}$$

The variables $\tilde{T}(\alpha,\beta)$ are also independent across (α,β) conditioned on the latent variables $\{\theta_i\}_{i\in[n]}$, and their variance is bounded above by the corresponding variances of $T(\alpha,\beta)$. Using the boundedness of $\tilde{T}(\alpha,\beta)$, by applying Bernstein's inequality with the choice of $z = 2\, n^{\psi}\left(\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|\right)^{1/2}$ for $\psi \in (0,\kappa)$, it follows that

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{\alpha<\beta\in S\times S} (\tilde{T}(\alpha,\beta) - \mathbb{E}[\tilde{T}(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}])\right| \right. \\
&\qquad \left. \ge z \;\middle|\; \mathcal{F}(u,v,t,x,\delta), \{\theta_i\}_{i\in[n]}\right) \\
&\le 2\exp\left(-\frac{\frac{z^2}{2}}{2\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|(1+o(1)) + \frac{2\phi^2 z}{3}}\right) \\
&= 2\exp(-n^{2\phi}(1-o(1))). \qquad \text{(VII.31)}
\end{aligned}$$

The last equality arises from the observation that t is chosen such that conditioned on $\mathcal{F}$, we can plug in (VII.30) to show that for our choice of z, it holds that

$$z = o(|S|^2 p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|).$$

In Lemma 7.9, we will show a bound on $\mathbb{P}\left(|T(\alpha,\beta)| \ge \phi^2\right)$, which translates to a bound on $\mathbb{E}[\mathbb{I}_{|T(\alpha,\beta)|\ge\phi^2}(|T(\alpha,\beta)| - \phi^2) \mid \mathcal{F}]$, which then upper bound the difference between the conditional expectations of T and $\tilde{T}$ according to

$$\begin{aligned}
&|\mathbb{E}[T(\alpha,\beta) \mid \mathcal{F}] - \mathbb{E}[\tilde{T}(\alpha,\beta) \mid \mathcal{F}]| \\
&\quad\le \mathbb{E}[|T(\alpha,\beta) - \tilde{T}(\alpha,\beta)| \mid \mathcal{F}] \\
&\quad= \mathbb{E}[|T(\alpha,\beta)| - \min(|T(\alpha,\beta)|, \phi^2) \mid \mathcal{F}] \\
&\quad= \mathbb{E}[\mathbb{I}_{|T(\alpha,\beta)|\ge\phi^2}(|T(\alpha,\beta)| - \phi^2) \mid \mathcal{F}]
\end{aligned}$$

Using this bound from Lemma 7.9 along with the conditions from $\mathcal{F}$ that guarantee $|S| = \Theta(n)$ and naively $|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}| = O(n)$, it follows that

$$\begin{aligned}
&\left|\sum_{\alpha<\beta\in S\times S} (\mathbb{E}[\tilde{T}(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}] - \mathbb{E}[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}])\right| \\
&\le (1+o(1))\binom{|S|}{2} \frac{2\phi}{\ln(|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|^{-1} p^{-1})} (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}| p)^{\phi} \\
&= O\left(|S|^2 (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}| p)^{\phi}\right) \\
&= O\left(n^2 \left(n^{-(\frac{1}{2}-\kappa)}\right)^{\phi}\right). \qquad \text{(VII.32)}
\end{aligned}$$

We choose $\phi = \lceil \frac{16}{1-2\kappa} \rceil \ge \frac{16}{1-2\kappa}$ so that this difference between the expectations of T and $\tilde{T}$ is $O(n^{-6})$.

By plugging in our choice of ϕ into Lemma 7.9, it also follows that

$$\mathbb{P}\left(\cup_{\alpha,\beta}\{\tilde{T}(\alpha,\beta) \ne T(\alpha,\beta)\} \mid \mathcal{F}\right) \le O(n^{-6}). \qquad \text{(VII.33)}$$

By combining (VII.31), (VII.32), and (VII.33), with probability at least $1 - 2\exp(-n^{2\psi}(1-o(1))) - O(n^{-6})$,

$$\begin{aligned}
&\left|\sum_{\alpha<\beta\in S\times S} \left(T(\alpha,\beta) - \mathbb{E}\left[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}\right]\right)\right| \\
&\qquad\le 2\, n^{\psi}\left(\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|\right)^{1/2} + O(n^{-6}), \qquad \text{(VII.34)}
\end{aligned}$$

where the first term will dominate the second term.

Finally we want to show concentration of the following expression with respect to the latent variables,

$$\frac{1}{\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \mathbb{E}\left[\sum_{\alpha<\beta\in S\times S} T(\alpha,\beta) \;\middle|\; \mathcal{F}, \{\theta_i\}_{i\in[n]}\right].$$

The expression can be written as a pairwise U-statistic,

$$U = \frac{1}{\binom{|S|}{2}} \sum_{\alpha<\beta\in S\times S} g(\theta_\alpha, \theta_\beta),$$

where g is a symmetric function, and

$$\begin{aligned}
&g(\theta_\alpha,\theta_\beta) \\
&= \frac{1}{p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \mathbb{E}\left[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}\right] \\
&= \frac{1}{p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{a\ne b\in[n]} N_{u,t}(a) N_{v,t}(b) \\
&\quad\times \mathbb{E}\left[M_B(a,(\alpha,\beta)) M_B(b,(\alpha,\beta)) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}\right] \\
&= \frac{1}{|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{a\ne b\in[n]} N_{u,t}(a) N_{v,t}(b) F(a,\alpha,\beta) F(b,\alpha,\beta).
\end{aligned}$$

It follows by boundedness of entries in F and the fact that $\|N_{u,t}\|_\infty \le 1$ and $\|N_{u,t}\|_0 = |\mathcal{S}_{u,t}|$, that $|g(\theta_\alpha,\theta_\beta)| \le 1$ almost surely. Therefore, by Lemma A.3 and choosing $z = \sqrt{8} n^{\psi} |S|^{-1/2}$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\binom{|S|}{2} p^2 |\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{\alpha<\beta\in S\times S} \left(\mathbb{E}\left[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}\right]\right.\right. \right. \\
&\qquad \left. \left. \left. - \mathbb{E}\left[T(\alpha,\beta) \mid \mathcal{F}\right]\right)\right| \ge z\right) \\
&\quad\le 2\exp\left(-\frac{|S| z^2}{8}\right) = 2\exp(n^{2\psi}). \qquad \text{(VII.35)}
\end{aligned}$$

The expected value with respect to the randomness in the latent variables is

$$\begin{aligned}
&\frac{1}{\binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\mathbb{E}\left[\sum_{\alpha<\beta\in S\times S} T(\alpha,\beta)\,\middle|\,\mathcal{F}\right]\\
&= \frac{1}{\binom{|S|}{2}}\sum_{\alpha<\beta\in S\times S}\frac{1}{|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{a\neq b\in[n]} N_{u,t}(a)N_{v,t}(b)\\
&\quad\times\mathbb{E}[F(a,\alpha,\beta)F(b,\alpha,\beta)]\\
&= \frac{1}{\binom{|S|}{2}}\sum_{\alpha<\beta\in S\times S}\frac{1}{|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\\
&\quad\times\sum_{a\neq b\in[n]} N_{u,t}(a)N_{v,t}(b)\sum_k \lambda_k^2 q_k(\theta_a)q_k(\theta_b)\\
&= \tilde{N}_{u,t}^T Q^T\Lambda^2 Q\tilde{N}_{v,t} - \sum_{a\in[n]}\tilde{N}_{u,t}(a)\tilde{N}_{v,t}(a)\sum_k\lambda_k^2 q_k^2(\theta_a).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\left|\sum_{a\in[n]}\tilde{N}_{u,t}(a)\tilde{N}_{v,t}(a)\sum_k\lambda_k^2 q_k^2(\theta_a)\right| &\le \frac{B^2\left(\sum_k\lambda_k^2\right)}{\max(|\mathcal{S}_{u,t}|,|\mathcal{S}_{v,t}|)}\\
&= O((|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{-1/2}).
\end{aligned}\tag{VII.36}$$

By combining (VII.34), (VII.35), and (VII.36), it follows that conditioned on $\mathcal{F}(u,v,s,\ell,x,\delta)$, with probability

$$1-2\exp\left(-n^{2\psi}(1-o(1))\right)-2\exp(n^{2\psi})-O(n^{-6}),$$

it holds that

$$\begin{aligned}
&\left|\frac{1}{\binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{\alpha<\beta\in S\times S}T(\alpha,\beta)-\tilde{N}_{u,t}^TQ^T\Lambda^2Q\tilde{N}_{v,t}\right|\\
&\le 2\,n^{\psi}\left(\binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|\right)^{-1/2}+O(n^{-6})+\frac{\sqrt{8}n^{\psi}}{|S|^{1/2}}\\
&\quad+o\left(\left(\binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|\right)^{-1}\right)+\frac{B^2\left(\sum_k\lambda_k^2\right)}{\max(|\mathcal{S}_{u,t}|,|\mathcal{S}_{v,t}|)}\\
&\le O\left(\frac{n^{\psi}}{(|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}}\right)+O\left(\frac{n^{\psi}}{|S|^{1/2}}\right)\\
&\quad+O\left(\frac{1}{(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}}\right).
\end{aligned}$$

Note that the third term is dominated by the first term as $|S|p=o(1)$. This completes the proof of Lemma 7.7. □

Lemma 7.8: Let $\mathcal{F}=\mathcal{F}(u,v,t,x,\delta)$ denote all the information revealed such that event $A'_{u,v,t}(x,\delta)$ holds.

$$\text{Var}[T(\alpha,\beta)\mid\mathcal{F},\{\theta_i\}_{i\in[n]}]\le 2\,p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|(1+o(1)).$$

Proof: To compute the variance of $T(\alpha,\beta)$ conditioned on $\mathcal{F},\{\theta_i\}_{i\in[n]}$, note that there is correlation in the terms within the sum of $T(\alpha,\beta)$ as there may be pairs (a,b) and (a',b') that share coordinates. In particular because the observation noise and sampling randomness for $M_B(a,b,c)$ is independent across different entries (a,b,c), then conditioned on $\{\theta_i\}_{i\in[n]}$, for $a\neq b$ and $a'\neq b'$, if all four coordinates $\{a,b,a',b'\}$ are distinct,

$$\text{Cov}[M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta)),M_B(a',(\alpha,\beta))M_B(b',(\alpha,\beta))]=0;$$

if $|\{a,b\}\cap\{a',b'\}|=2$, i.e. $(a',b')=(a,b)$ or $(a',b')=(b,a)$,

$$\begin{aligned}
&\left|\text{Cov}[M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta)),M_B(a',(\alpha,\beta))M_B(b',(\alpha,\beta))]\right|\\
&=\text{Var}[M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta))]\\
&\le\mathbb{E}[{M_B}^2(a,(\alpha,\beta)){M_B}^2(b,(\alpha,\beta))]\le p^2;
\end{aligned}$$

and if $\{a,b\}\cup\{a',b'\}=\{x,y,z\}$ such that $\{a,b\}\cap\{a',b'\}=\{x\}$, then

$$\begin{aligned}
&\Big|\text{Cov}[M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta)),M_B(a',(\alpha,\beta))\\
&\quad\times M_B(b',(\alpha,\beta))\mid\mathcal{F},\{\theta_i\}_{i\in[n]}]\Big|\\
&=|\text{Var}[M_B(x,(\alpha,\beta))]\mathbb{E}[M_B(y,(\alpha,\beta))]\mathbb{E}[M_B(z,(\alpha,\beta))]|\\
&\le\left|\mathbb{E}[{M_B}^2(x,(\alpha,\beta))]\mathbb{E}[M_B(y,(\alpha,\beta))]\mathbb{E}[M_B(z,(\alpha,\beta))]\right|\\
&\le p^3.
\end{aligned}$$

The inequalities follow from the property that every entry of M_B has absolute value bounded by 1, and takes value 0 with probability $(1-p)$ in the event it is not observed.

We use this to expand the variance calculation, and use the properties that for every entry a, $|N_{u,t}(a)|\le\mathbb{I}_{(a\in\mathcal{S}_{u,t})}$. We have dropped the conditioning notation due to the length of the expressions.

The first term in (VII.37) dominates the latter terms because $p|\mathcal{S}_{u,t}|\le pn=o(1)$ and $p|\mathcal{S}_{v,t}|=o(1)$. The final variance calculation that uses the above inequalities to prove the lemma is given in (VII.37) shown at the bottom of the next page.

Lemma 7.9:

$$\mathbb{P}\left(|T(\alpha,\beta)|\ge z\mid\mathcal{F}\right)\le(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\lceil\sqrt{z}\rceil}(1+o(1).$$

As a result,

$$\mathbb{P}\left(\cup_{\alpha,\beta}\{|T(\alpha,\beta)|\ge z\}\mid\mathcal{F}\right)\le\binom{|S|}{2}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\lceil\sqrt{z}\rceil}(1+o(1)),$$

and

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}_{|T(\alpha,\beta)|\ge\phi^2}(|T(\alpha,\beta)|-\phi^2)\mid\mathcal{F}]\\
&\quad\le(1+o(1))\frac{2\phi}{\ln(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|^{-1}p^{-1})}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\phi}
\end{aligned}$$

Proof: Let us define

$$\begin{aligned}
Z_u(\alpha,\beta)&=\{a\in[n]\ s.t.\ a\in\mathcal{S}_{u,t},(a,\alpha,\beta)\in\Omega_1\},\\
Z_v(\alpha,\beta)&=\{b\in[n]\ s.t.\ b\in\mathcal{S}_{v,t},(b,\alpha,\beta)\in\Omega_1\}.
\end{aligned}$$

Furthermore, because $|M_B(\cdot,\cdot,\cdot)|\le1$ and $\|N_{u,s}\|_\infty\le1$, for any $a,b\in[n]$, it follows that

$$\begin{aligned}
&|N_{u,t}(a)N_{v,t}(b)M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta))|\\
&\quad\le\mathbb{I}_{(a\in Z_u(\alpha,\beta))}\mathbb{I}_{(b\in Z_v(\alpha,\beta))},
\end{aligned}$$

which implies

$$|T(\alpha,\beta)|\le|Z_u(\alpha,\beta)||Z_v(\alpha,\beta)|\le|Z_u(\alpha,\beta)\cup Z_v(\alpha,\beta)|^2.$$

Note that $|Z_u(\alpha,\beta) \cup Z_v(\alpha,\beta)| \sim \text{Binomial}(|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|, p)$. It follows then that

$$\begin{aligned}
&\mathbb{P}\left(|T(\alpha,\beta)| \geq z \mid \mathcal{F}\right)\\
&\leq \mathbb{P}\left(|Z_u(\alpha,\beta) \cup Z_v(\alpha,\beta)|^2 \geq z \mid \mathcal{F}\right)\\
&= \mathbb{P}\left(|Z_u(\alpha,\beta) \cup Z_v(\alpha,\beta)| \geq \lceil\sqrt{z}\rceil \mid \mathcal{F}\right)\\
&= \sum_{i=\lceil\sqrt{z}\rceil}^{|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|} \binom{|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|}{i} p^i (1-p)^{|\mathcal{S}_{u,t}\cap\mathcal{S}_{v,t}|-i}\\
&\leq (1-p)^{|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|} \sum_{i=\lceil\sqrt{z}\rceil}^{|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|} \left(\frac{|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p}{1-p}\right)^i\\
&\leq (1-p)^{|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|} \left(\frac{|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p}{1-p}\right)^{\lceil\sqrt{z}\rceil} \sum_{i=0}^{\infty} \left(\frac{|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p}{1-p}\right)^i\\
&\leq (1-p)^{|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|} \left(\frac{|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p}{1-p}\right)^{\lceil\sqrt{z}\rceil} (1+o(1))\\
&\leq (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p)^{\lceil\sqrt{z}\rceil} (1+o(1))\\
&\leq (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p)^{\sqrt{z}} (1+o(1)),
\end{aligned}$$

where we used the fact that $|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p = o(1)$.

We use the bound on the tail probabilities to show that

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}_{|T(\alpha,\beta)|\geq\phi^2}(|T(\alpha,\beta)| - \phi^2) \mid \mathcal{F}]\\
&= \int_0^\infty \mathbb{P}\left(|T(\alpha,\beta)| \geq \phi^2 + z\right) dz\\
&\leq (1+o(1)) \int_0^\infty (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p)^{\sqrt{\phi^2+z}}\, dz\\
&\leq (1+o(1)) \int_\phi^\infty 2\ y\, (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p)^y\, dy\\
&= (1+o(1)) \frac{2\phi}{\ln(|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|^{-1}p^{-1})} (|\mathcal{S}_{u,t} \cup \mathcal{S}_{v,t}|p)^\phi.
\end{aligned}$$

□

E. Proof of Lemma 6.2

Proof: Now we are ready to bound the difference between $d(u,v)$ and $\hat{d}(u,v)$ for any $u,v \in [n]$. Recall,

$$\begin{aligned}
d(\theta_u,\theta_v) &= \|\Lambda^{2t+1}Q(e_u - e_v)\|^2\\
&= (e_u - e_v)^T Q^T \Lambda^{4t+2} Q (e_u - e_v)\\
&= e_u^T Q^T \Lambda^{4t+2} Q e_u + e_v^T Q^T \Lambda^{4t+2} Q e_v\\
&\quad - e_u^T Q^T \Lambda^{4t+2} Q e_v - e_v^T Q^T \Lambda^{4t+2} Q e_u, \quad \text{(VII.38)}
\end{aligned}$$

and according to (IV.5),

$$\mathsf{dist}(u,v) = \frac{1}{\binom{|S|}{2}p^2}(Z_{uu} + Z_{vv} - Z_{uv} - Z_{vu}) \quad \text{(VII.39)}$$

for $S \equiv S_{u,s,t} = n \setminus (\mathcal{B}_{u,t} \cup \mathcal{B}_{v,t} \cup [n/2])$ and

$$Z_{uv} = \frac{1}{\binom{|S_{u,s,t}|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{\alpha<\beta\in S_{u,s,t}\times S_{u,s,t}} T_{uv}(\alpha,\beta). \quad \text{(VII.40)}$$

By Lemma 7.1, event $\mathcal{A}^3_{u,t}(\delta)$ holds with probability at least $1 - O\Big(n\exp\big(-\Theta(n^{2\kappa})\big)\Big)$. By Lemmas 7.2 and 7.6, conditioned on $\mathcal{A}^3_{u,t}(\delta)$, for

$$x = \frac{16\lambda_{\max}^{2t-2}n^\psi}{(1-\delta)n^\kappa} = o(1),$$

event $A'_{u,v,t}(x,\delta)$ holds with probability at least $1 - 4\,r\exp(-n^{2\psi}(1-o(1)))$, implying

$$\begin{aligned}
&|\tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t} - e_u^T Q^T \Lambda^{2(2t+1)} Q e_v|\\
&\quad \leq \frac{n^\psi}{n^\kappa}\left(16B\lambda_{\max}^{2(t-1)}\left(\sum_{k=1}^r 2\lambda_k^{2(t+1)}\right)(1-\delta)^{-1}(1+o(1))\right).
\end{aligned}$$

By Lemma 7.7, conditioned on $A'_{u,v,t}(x,\delta)$, with probability $1 - 4\exp(-n^{2\psi}(1-o(1)))$,

$$\begin{aligned}
|Z_{uv} - \tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t}| &= O\left(\frac{n^\psi}{(|S_{u,s,t}|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}}\right)\\
&\quad + O\left(\frac{n^\psi}{|S_{u,s,t}|^{1/2}}\right),
\end{aligned}$$

where $|S_{u,s,t}| = \Theta(n) = \Omega(n^{2\kappa})$, by event $A'_{u,v,t}(x,\delta)$ and $t \geq \frac{1}{4\kappa}$,

$$|S_{u,s,t}|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}| = \Theta(n^2n^{-3+2\kappa}n^{4\kappa t}) = \Omega(n^{2\kappa}).$$

$$\begin{aligned}
&\text{Var}[T(\alpha,\beta) \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}]\\
&= \sum_{a\neq b\in[n]} \Big(N_{u,t}^2(a)N_{v,t}^2(b) + N_{u,t}(a)N_{v,t}(b)N_{u,t}(b)N_{v,t}(a)\Big)\text{Var}[M_B(a,(\alpha,\beta))M_B(b,(\alpha,\beta))]\\
&\quad + \sum_{a\neq b\in[n]}\sum_{c\notin\{a,b\}} \Big(N_{u,t}^2(a)N_{v,t}(b)N_{v,t}(c) + N_{v,t}^2(a)N_{u,t}(b)N_{u,t}(c) + N_{u,t}(a)N_{u,t}(b)N_{v,t}(a)N_{v,t}(c)\\
&\qquad\qquad + N_{u,t}(a)N_{u,t}(c)N_{v,t}(a)N_{v,t}(b)\Big)\text{Var}[M_B(a,(\alpha,\beta))]\mathbb{E}[M_B(b,(\alpha,\beta))]\mathbb{E}[M_B(c,(\alpha,\beta))]\\
&\leq p^2 \sum_{a\neq b\in[n]} (\mathbb{I}_{(a\in\mathcal{S}_{u,t},b\in\mathcal{S}_{v,t})} + \mathbb{I}_{(\{a,b\}\subset\mathcal{S}_{u,t}\cap\mathcal{S}_{v,t})})\\
&\quad + p^3 \sum_{a\neq b\in[n]}\sum_{c\notin\{a,b\}} \Big(\mathbb{I}_{(a\in\mathcal{S}_{u,t},\{b,c\}\subset\mathcal{S}_{v,t})} + \mathbb{I}_{(a\in\mathcal{S}_{v,t},\{b,c\}\subset\mathcal{S}_{u,t})} + \mathbb{I}_{(a\in\mathcal{S}_{u,t}\cap\mathcal{S}_{v,t},b\in\mathcal{S}_{u,t},c\in\mathcal{S}_{v,t})} + \mathbb{I}_{(a\in\mathcal{S}_{u,t}\cap\mathcal{S}_{v,t},c\in\mathcal{S}_{u,t},b\in\mathcal{S}_{v,t})}\Big)\\
&\leq 2\ p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}| + 2\ p^3|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|^2 + 2\ p^3|\mathcal{S}_{u,t}|^2|\mathcal{S}_{v,t}|\\
&= 2\ p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|(1+o(1)). \quad \text{(VII.37)}
\end{aligned}$$

To put it all together, for $\psi \in (0, \kappa)$, with probability at least

$$1 - O\Big(\exp(-n^{2\psi}(1 - o(1)))\Big),$$

it holds that

$$|\mathsf{dist}(u, v) - d(\theta_u, \theta_v)| = O\left(\frac{r\lambda_{\max}^{4t} n^{\psi}}{n^{\kappa}}\right).$$

This completes the proof of Lemma 6.2. □

VIII. Proof of Lemma 6.3: Perturbation Analysis of Distance

We establish the proof of Lemma 6.3 here. To do so, we establish a perturbation property of dist here, which combined with Lemma 6.2 will result into the proof of Lemma 6.3.

We study the perturbation in the dist estimate when each noisy observed entry is arbitrarily perturbed. Specifically, for any $(u, v, w) \in [n]^3$, $M_1(u, v, w)$ is observed with probability p. If observed, according to (III.1), $M_1(u, v, w) = F(u, v, w) + \epsilon_{uvw} = F_r(u, v, w) + \varepsilon_{uvw} + \epsilon_{uvw}$, where F_r is the best rank r approximation to F. This expression shows that we can interpret the deviation from a rank r model as a deterministic perturbation of ε_{uvw}, bounded in absolute value by ε. Note that ε_{uvw} can be any arbitrary (or adversarial), unknown deterministic quantity satisfying $|\varepsilon_{uvw}| \le \varepsilon$.

Lemma 8.1 provides a bound on the perturbation in the distance estimate, dist, that results from these entrywise perturbations of the observations.

Lemma 8.1: Let $p = n^{-3/2+\kappa}$ for $\kappa \in (0, \frac{1}{2})$, t as defined in (IV.3), $\delta \in (0, \frac{1}{2})$, and

$$x = \frac{16\lambda_{\max}^{2t-2} n^{\psi}}{(1-\delta)n^{\kappa}}.$$

For any $u \in [n]$, recall the event

$$A'_{u,v,t}(x, \delta) = \cap_{k=1}^{r}\big(\mathcal{A}^4_{u,k,t}(x, \delta) \cap \mathcal{A}^4_{v,k,t}(x, \delta)\big). \quad \text{(VIII.1)}$$

Let event $A'_{u,v,t}(x, \delta)$ hold. Let each observed entry of M_1 be perturbed by adding arbitrary, deterministic quantity bounded by $\varepsilon \ge 0$. Then for any $u, v \in [n]^2$, the distance estimate $\mathsf{dist}(u, v)$ is perturbed by at most $O(t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2})$ with probability at least $1 - \exp\big(-\Omega(n^{2\kappa})\big) - O(n^{-8})$.

Proof: Recall definition of dist in (IV.5):

$$\mathsf{dist}(u, v) = (Z_{uu} + Z_{vv} - Z_{uv} - Z_{vu}),$$

$$Z_{uv} = \frac{1}{|\mathcal{V}_B(u, v, t)|p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \sum_{(\alpha,\beta)\in\mathcal{V}_B(u,v,t)} T_{uv}(\alpha, \beta),$$

$$\mathcal{V}_B(u, v, t) = \{(\alpha, \beta) \in \mathcal{V}_B \ s.t. \ \alpha \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}, \beta \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}\},$$

$$T_{uv}(\alpha, \beta) = \sum_{a\neq b\in[n]} N_{u,t}(a)N_{v,t}(b)M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta))$$

We shall bound the perturbation on Z_{uv}. Similar bounds will follow for the other three terms which will conclude the main results. Our interest is in understanding how does Z_{uv} change if each observed entry is changed by arbitrary quantity bounded by $\varepsilon \ge 0$. This will induce a bound on the changes in $T_{uv}(\cdot, \cdot)$ which will help bound the change in Z_{uv}. By assumption (VIII.1), $A'_{u,v,t}(x, \delta)$ holds. Conditioned on event $A'_{u,v,t}(x, \delta)$, all the information associated with $\mathcal{B}_{u,2t}$ and $\mathcal{B}_{v,2t}$ is revealed; however, information about $[n]\backslash(\mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t})$ is not. Let $\mathcal{F}(u, v, t, x, \delta)$ denote all the information revealed such that event $A'_{u,v,t}(x, \delta)$ holds.

Under $A'_{u,v,t}(x, \delta)$, by definition $\mathcal{A}^3_{u,t}(\delta)$ and $\mathcal{A}^3_{v,t}(\delta)$ holds. This implies that for $S \equiv \mathcal{V}_B(u, v, t)$, $|S| = \frac{n}{2} - o(n) = \frac{n(1-o(1))}{2}$. Furthermore, it implies that $|\mathcal{S}_{u,t}|$ and $|\mathcal{S}_{v,t}|$ are both greater than or equal to $(1-\delta)^{2t}2^{-3t-1}n^{2\kappa t}(1-o(1))$. As shown in (VII.30),

$$|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}| = \Omega(n^{2\kappa}) \quad \text{(VIII.2)}$$

results from the choice of $t \ge \frac{1}{4\kappa}$, and the fact that δ and t are both constants.

For given $\alpha \neq \beta \in \mathcal{V}_B(u, v, t)$, $T_{uv}(\alpha, \beta)$ is summation over terms, indexed by $a \neq b \in [n]$, containing product $N_{u,t}(a)N_{v,t}(b)M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta))$. Now $N_{u,t}(a) = 0$ if $a \notin \mathcal{S}_{u,t}$, $N_{v,t}(b) = 0$ if $b \notin \mathcal{S}_{v,t}$. For $a \in \mathcal{S}_{u,t}$, $N_{u,t}(a)$ is product of $2t$ terms, each bounded in absolute value by 1: let $N_{u,t}(a) = \prod_{i=1}^{2t} w_i$ with $|w_i| \le 1$ for all $i \le 2t$. Let ε_i be arbitrary, deterministic quantity added to w_i with $|\varepsilon_i| \le \varepsilon$ for $i \le 2t$. Then change in $N_{u,t}(a)$ is bounded as

$$\begin{aligned}
&\Big|\prod_{i=1}^{2t} w_i - \prod_{i=1}^{2t}(w_i + \varepsilon_i)\Big| \\
&= \Big|\sum_{S\subset[2t]:S\neq\emptyset}\prod_{i\in S}\varepsilon_i \prod_{s\in[2t]\backslash S} w_i\Big| \\
&\le \sum_{S\subset[2t]:S\neq\emptyset}\prod_{i\in S}|\varepsilon_i| \prod_{s\in[2t]\backslash S}|w_i| \\
&\le \sum_{S\subset[2t]:S\neq\emptyset}\varepsilon^{|S|} = \sum_{i=1}^{2t}\binom{2t}{i}\varepsilon^s \\
&= \varepsilon\Big(\sum_{i=0}^{2t-1}\frac{(2t)!}{(2t-i-1)!(i+1)!}\varepsilon^i\Big) \\
&\le 2t\varepsilon\Big(\sum_{i=0}^{2t-1}\frac{(2t-1)!}{((2t-1)-i)!i!}\varepsilon^i\Big) \\
&= 2t\varepsilon\Big(\sum_{i=0}^{2t-1}\binom{2t-1}{i}\varepsilon^i\Big) \\
&= 2t\varepsilon(1+\varepsilon)^{2t-1} \equiv \Delta(t, \varepsilon). \quad \text{(VIII.3)}
\end{aligned}$$

That is, $N_{u,t}(a)$ changes by at most $\Delta(t, \varepsilon)$. Similarly $N_{v,t}(b)$ changes by at most $\Delta(t, \varepsilon)$. Therefore, $N_{u,t}(a)N_{v,t}(b)$ can change at most by $O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)$.

By definition $|M_B(a, (\alpha, \beta))|, |M_B(b, (\alpha, \beta))| \le 1$. Further, $M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta)) \neq 0$ only if $\mathbb{I}_{((a,\alpha,\beta)\in\Omega_1)}\mathbb{I}_{((b,\alpha,\beta)\in\Omega_1)} = 1$. Therefore, we can bound change in the term $N_{u,t}(a)N_{v,t}(b)M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta))$ as $\mathbb{I}_{((a,\alpha,\beta)\in\Omega_1)}\mathbb{I}_{((b,\alpha,\beta)\in\Omega_1)}O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)$. Therefore, we can bound the change in Z_{uv} by

$$\frac{O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)}{|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|} \Big(\sum_{a\in\mathcal{S}_{u,t},b\in\mathcal{S}_{v,t},\alpha,\beta\in S}\mathbb{I}_{((a,\alpha,\beta)\in\Omega_1)}\mathbb{I}_{((b,\alpha,\beta)\in\Omega_1)}\mathbb{I}_{a\neq b}\Big) \quad \text{(VIII.4)}$$

$$= \frac{O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)}{|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{\alpha,\beta\in S} X_{\alpha\beta} \quad \text{(VIII.5)}$$

where

$$X_{\alpha\beta} = \sum_{a\in\mathcal{S}_{u,t}, b\in\mathcal{S}_{v,t}, a\neq b} \mathbb{I}_{((a,\alpha,\beta)\in\Omega_1)}\mathbb{I}_{((b,\alpha,\beta)\in\Omega_1)}.$$

To conclude the Lemma, it will be sufficient to argue that $\sum_{\alpha,\beta\in S} X_{\alpha\beta} = O(|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)$ with high probability given $\mathcal{F}$. We use a similar argument as the proof of Lemma 7.7. Given $\mathcal{F} \equiv \mathcal{F}(u,v,t,x,\delta)$, $\{X_{\alpha\beta}\}_{\alpha,\beta\in S^2}$ are conditionally independent random variables. By the same argument as that in Lemma 7.9, it follows that

$$\mathbb{P}\left(\cup_{\alpha,\beta}\{X_{\alpha,\beta}\geq\phi^2\} \mid \mathcal{F}\right) \leq \binom{|S|}{2}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\phi}(1+o(1)) \tag{VIII.6}$$

$$\mathbb{E}[\mathbb{I}_{X_{\alpha,\beta}\geq\phi^2}(X_{\alpha,\beta}-\phi^2) \mid \mathcal{F}] \leq (1+o(1))\frac{2\phi}{\ln(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|^{-1}p^{-1})}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\phi}. \tag{VIII.7}$$

We define $\tilde{X}_{\alpha\beta} = \min(X_{\alpha\beta},\phi^2)$ for $\phi = \lceil 16/(1-2\kappa)\rceil$ so that $|\tilde{X}_{\alpha\beta} - \mathbb{E}[\tilde{X}_{\alpha\beta}]| \leq \phi^2$. By (VIII.6), and the choice of ϕ along with the conditions from $\mathcal{F}$ that guarantee $|S| = \Theta(n)$ and naively $|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}| = O(n)$,

$$\mathbb{P}\left(\sum_{\alpha,\beta}X_{\alpha,\beta} \neq \sum_{\alpha,\beta}\tilde{X}_{\alpha,\beta} \mid \mathcal{F}\right) \leq \binom{|S|}{2}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\phi}(1+o(1)) = O\big(n^{-6}\big). \tag{VIII.8}$$

By (VIII.7),

$$\begin{aligned}|\mathbb{E}[X_{\alpha,\beta} \mid \mathcal{F}] - \mathbb{E}[\tilde{X}_{\alpha,\beta} \mid \mathcal{F}]| &\leq \mathbb{E}[\mathbb{I}_{X_{\alpha,\beta}\geq\phi^2}(X_{\alpha,\beta}-\phi^2) \mid \mathcal{F}] \\ &\leq (1+o(1))\frac{2\phi}{\ln(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|^{-1}p^{-1})}(|\mathcal{S}_{u,t}\cup\mathcal{S}_{v,t}|p)^{\phi} \\ &= O\big(n^{-6}\big).\end{aligned} \tag{VIII.9}$$

By the same argument as that in Lemma 7.8, it follows that

$$\begin{aligned}\mathrm{Var}[\tilde{X}_{\alpha,\beta} \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}] &\leq \mathrm{Var}[X_{\alpha,\beta} \mid \mathcal{F}, \{\theta_i\}_{i\in[n]}] \\ &\leq 2\ p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|(1+o(1)).\end{aligned}$$

By Bernstein's inequality, for $z = 2\ n^{\psi}|S|p(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}$ for $\psi\in(0,\kappa)$,

$$\begin{aligned}\mathbb{P}\left(\sum_{\alpha,\beta\in S}\big(\tilde{X}_{\alpha,\beta} - \mathbb{E}[\tilde{X}_{\alpha,\beta}]\big) > z\right) &\leq \exp\Big(-\frac{3z^2}{2z\phi^2 + 12(1+o(1))|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\Big) \\ &= \exp(-n^{2\psi}(1-o(1))).\end{aligned} \tag{VIII.10}$$

By (VIII.5), (VIII.8), (VIII.9), and (VIII.10), given $A'_{u,v,t}(x,\delta)$ holds, with probability $1-\exp\big(-n^{2\kappa}(1-o(1))\big) - O(n^{-6})$, the change in Z_{uv} is bounded above by

$$\begin{aligned}&\frac{O(\Delta(t,\varepsilon)+\Delta(t,\varepsilon)^2)}{|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{\alpha,\beta\in S}X_{\alpha\beta} \\ &= \frac{O(\Delta(t,\varepsilon)+\Delta(t,\varepsilon)^2)}{|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\sum_{\alpha,\beta\in S}\tilde{X}_{\alpha\beta} \\ &= \frac{O(\Delta(t,\varepsilon)+\Delta(t,\varepsilon)^2)}{|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|}\Big(\mathbb{E}[\sum_{\alpha,\beta\in S}X_{\alpha\beta} \mid \mathcal{F}] + 2\,n^{\psi}|S|p(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2}\Big).\end{aligned}$$

By (VIII.2), this choice of $2n^{\psi}|S|p(|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)^{1/2} = o(|S|^2p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|)$ for $\psi\in(0,\kappa)$. Finally we use the bound that $\mathbb{E}[\sum_{\alpha,\beta\in S}X_{\alpha\beta} \mid \mathcal{F}] = \binom{|S|}{2}p^2|\mathcal{S}_{u,t}||\mathcal{S}_{v,t}|$ to argue that with high probability the change in Z_{uv} is bounded above by

$$O(\Delta(t,\varepsilon)+\Delta(t,\varepsilon)^2) = O(t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}).$$

This completes the proof of Lemma 8.1.

□

A. Completing Proof of Lemma 6.3

Under the setup of Lemma 6.3, as argued in the proof of Lemma 6.2, $A'_{u,v,t}(x,\delta)$, with appropriate choice of x,δ as considered in statement of Lemma 8.1, holds with probability at least $1 - 4\ r\exp(-n^{2\psi}(1-o(1)))$. And dist (without perturbation), is within $O\left(n^{-(\kappa-\psi)}\right)$ for any pair of $u,v\in[n]$. By Lemma 8.1, under event $A'_{u,v,t}(x,\delta)$, the dist is further perturbed by $O\left(t\varepsilon(1+\varepsilon)^{2t-1} + t^2\varepsilon^2(1+\varepsilon)^{4t-2}\right)$ with probability at least $1 - O\Big(\exp(-n^{2\psi}(1-o(1)))\Big) - O\Big(n^{-6}\Big)$. Putting these together, we conclude the claim of Lemma 6.3.

APPENDIX
USEFUL LEMMAS AND OMITTED PROOFS

We present the Proof of Lemma 6.1 below.

Proof: [Lemma 6.1] We assumed the algorithm has access to two fresh samples, where M_1 is used to compute $\hat{d}$, and M_2 is used to compute the final estimate $\hat{F}$. Alternatively one could effectively obtain two sample sets by sample splitting. For some $(a,b,c)\in\Omega_2$, the observation $M_2(a,b,c)$ is independent of $\hat{d}$, and $\mathbb{E}[M_2(a,b,c)] = f(\theta_a,\theta_b,\theta_c)$. Conditioned on Ω_2, by definition of $\hat{F}$ and by assuming properties 6.1 and 6.2, it follows that

$$\begin{aligned}&\mathbb{E}[(\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w))^2] \\ &= \left(\frac{1}{|\Omega_{2uvw}|}\sum_{(a,b,c)\in\Omega_{2uvw}} f(\theta_a,\theta_b,\theta_c) - f(\theta_u,\theta_v,\theta_w)\right)^2 \\ &\quad + \frac{1}{|\Omega_{2uvw}|^2}\sum_{(a,b,c)\in\Omega_{2uvw}}\mathrm{Var}[M_2(a,b,c)] \\ &\overset{(a)}{\leq} \mathrm{bias}^2(\eta+\Delta) + \frac{\sigma^2}{|\Omega_{2uvw}|}.\end{aligned}$$

Inequality (a) follows from property 6.1 and property 6.2 for all $3n$ tuples $\{(u,a): a\in[n]\}\cup\{(v,b): b\in[n]\}\cup\{(w,c): c\in[n]\}$: $|d(u,a)-\hat{d}(u,a)|\leq\Delta$ and $\hat{d}(u,a)\leq\eta \implies d(u,a)\leq\eta+\Delta$, similarly $d(v,b), d(w,c)\leq\eta+\Delta$. As per (III.1), we have that $\mathrm{Var}[M_2(a,b,c)]\leq\sigma^2$ for all $(a,b,c)\in\Omega_2$. Define $\mathcal{V}_{uvw} = \{(a,b,c)\in[n]^3 : d(u,a)<\eta-\Delta,\ d(v,b)<\eta-\Delta,\ d(w,c)<\eta-\Delta\}$. Assuming property 6.3,

$$\begin{aligned}|\mathcal{V}_{uvw}| &= |\{a\in[n]: d(u,a)<\eta-\Delta\}|\ |\{b\in[n]: d(v,b)<\eta-\Delta\}| \\ &\quad\times|\{c\in[n]: d(w,c)<\eta-\Delta\}| \\ &\geq (\mathsf{meas}(\eta-\Delta)n)^3.\end{aligned}$$

By the Bernoulli sampling model, each tuple $(a,b,c)\in[n]^3$ belongs to Ω_2 with probability p independently. By a

straightforward application of Chernoff's bound, it follows that for any $\delta \in (0,1)$,

$$\mathbb{P}\left(|\Omega_2 \cap \mathcal{V}_{uvc}| \leq (1-\delta)\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right) \leq \exp\left(-\frac{\delta^2 p\left(\mathsf{meas}(\eta-\Delta)n\right)^3}{2}\right). \tag{A.1}$$

Therefore, by assuming property 6.2 for $3n$ tuples $\{(u,a) : a \in [n]\} \cup \{(v,b) : b \in [n]\} \cup \{(w,c) : c \in [n]\}$, it follows that with probability at least $1 - \exp\left(-\frac{\delta^2 p(\mathsf{meas}(\eta-\Delta)n)^3}{2}\right)$,

$$\begin{aligned}|\Omega_{2uvw}| &= |\{(a,b,c) \in \Omega_2 \ : \hat{d}(u,a) < \eta,\ \hat{d}(v,b) < \eta,\ \hat{d}(w,c) < \eta\}| \\ &\geq |\{(a,b,c) \in \Omega_2 \ : d(u,a) < \eta - \Delta, \\ &\qquad \mathrm{d(v,\ b)}\ < \eta - \Delta,\ d(w,c) < \eta - \Delta\}| \\ &= |\Omega_2 \cap \mathcal{V}_{uvw}| \\ &\geq (1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3.\end{aligned}$$

Define the event $\mathcal{H} = \{|\Omega_{2uvw}| \geq (1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3|\}$. It follows that $\mathbb{P}(\mathcal{H}^c) \leq \exp\left(-\frac{1}{2}\delta^2 p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right)$. By definition, $F(u,v,w) = f(\theta_u,\theta_v,\theta_w) \in [0,1]$ for all $u,v,w \in [n]$. Therefore,

$$\begin{aligned}&\mathbb{E}[(\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w))^2] \\ &\leq \mathbb{E}[(\hat{F}(u,v,w) - f(\theta_u,\theta_v,\theta_w))^2 \,\Big|\, \mathcal{H}] + \mathbb{P}(\mathcal{H}^c) \\ &\leq \mathsf{bias}^2(\eta+\Delta) + \frac{1}{(1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3} \\ &\quad + \exp\left(-\frac{1}{2}\delta^2 p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right).\end{aligned}$$

We add an additional $3n\alpha_1 + \alpha_2$ in the final MSE bound: $3n\alpha_1$ for violation of property 6.2 for any of the $3n$ tuples $\{(u,a) : a \in [n]\} \cup \{(v,b) : b \in [n]\} \cup \{(w,c) : c \in [n]\}$, and α_2 for violation of property 6.3.

To obtain the high-probability bound, note that $M_2(a,b,c)$ are independent across indices $(a,b,c) \in \Omega_2$ as well as independent of observations in Ω_1. Additionally, the model assumes that $F(a,b,c), M_2(a,b,c) \in [0,1]$, and $\mathbb{E}[M_2(a,b,c)] = F(a,b,c)$ for observed tuples (a,b,c). By an application of Hoeffding's inequality for bounded, zero-mean independent variables, for any $\delta' \in (0,1)$ it follows that assuming property 6.1, property 6.2 for $3n$ tuples $\{(u,a) : a \in [n]\} \cup \{(v,b) : b \in [n]\} \cup \{(w,c) : c \in [n]\}$, and property 6.3 hold, we have

$$\mathbb{P}\left(\frac{\left|\sum_{(a,b,c)\in\Omega_{2uvw}}(M(a,b,c)-F(a,b,c))\right|}{|\Omega_{2uvw}|} \geq \delta' \,\Big|\, \mathcal{H}\right) \leq \exp\left(-\delta'^2(1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right).$$

Therefore,

$$|\hat{F}_{uvw} - f(\theta_u,\theta_v,\theta_w)| \leq \mathsf{bias}(\eta+\Delta) + \delta',$$

with probability at least

$$1 - \exp\left(-\frac{1}{2}\delta^2 p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right) - \exp\left(-\delta'^2(1-\delta)p\left(\mathsf{meas}(\eta-\Delta)n\right)^3\right) - 3n\alpha_1 - \alpha_2.$$

This completes the proof of Lemma 6.1. □

Lemma A.1: The following inequalities hold:

(a) For any $\rho \geq 2$ and integer $r \geq 1$,

$$\sum_{s=1}^{r} \rho^s \leq 2\rho^r.$$

(b) For any $\rho \geq 2$ and non-negative integer s,

$$\rho^s \geq s\rho.$$

(c) Further, if $\exp(-a\rho) \leq \frac{1}{2}$ for some $a > 0$, then

$$\sum_{s=1}^{r} \exp(-a\rho^s) \leq 2\exp(-a\rho)$$

Proof: To prove (a), note that for any $\rho \geq 2$,

$$\sum_{s=1}^{r} \rho^s \leq \rho^r \sum_{s=1}^{r} \rho^{s-r} = \rho^r \sum_{s=0}^{r-1} \rho^{-s} \leq \rho^r \sum_{s=0}^{r-1} 2^{-s} \leq 2\rho^r.$$

To prove (b), first check that it trivially holds for $s = 0$ and $s = 1$. The inequality holds for $s = 2$ iff $\rho \geq 2$. The inequality hold for s iff $\rho \geq s^{1/(s-1)}$. We can verify that $s^{1/(s-1)}$ is a decreasing function in s, such that if the inequality holds for $s = 2$, it will also hold for $s \geq 2$. To prove (c), further consider $\exp(-a\rho) \leq \frac{1}{2}$,

$$\begin{aligned}\sum_{s=1}^{r} \exp(-a\rho^s) &\leq \sum_{s=1}^{r} \exp(-as\rho) \\ &\leq \exp(-a\rho) \sum_{s=1}^{r} \exp(-a\rho(s-1)) \\ &\leq \exp(-a\rho) \sum_{s=0}^{r-1} \exp(-a\rho s) \\ &\leq \exp(-a\rho) \sum_{s=0}^{r-1} 2^{-s} \\ &\leq 2\exp(-a\rho).\end{aligned}$$

□

Lemma A.2: If $\mathbb{P}(|X| \geq z) \leq c\exp(-\frac{z^2}{Q})$, then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2\nu^2}{2}\right)$$

with $\nu = \sqrt{\frac{Q(1+c^2\pi)}{2}}$.

Proof:

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &= \int_0^\infty \mathbb{P}\left(e^{\lambda X} \geq Z\right) dZ \\ &= \int_{-\infty}^\infty \mathbb{P}\left(\lambda X \geq z\right) e^z dz \\ &\leq \int_{-\infty}^\infty c\exp(-\frac{z^2}{Q\lambda^2} + z) dz \\ &\leq c\exp(\frac{\lambda^2 Q}{4}) \int_{-\infty}^\infty \exp(-\frac{1}{\lambda^2 Q}(z - \frac{\lambda^2 Q}{2})^2) dz \\ &\leq c\exp(\frac{Q\lambda^2}{4})\sqrt{\pi\lambda^2 Q}.\end{aligned}$$

Using the fact that $\sqrt{x} < e^{x/5}$ for all $x \geq 0$, it follows that

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\frac{Q\lambda^2}{4} + \frac{c^2\pi\lambda^2 Q}{4})$$

Therefore, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\frac{(1 + c^2\pi)Q\lambda^2}{4}).$$

□

Lemma A.3: Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Let $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function. Consider U-statistics with respect to g of $X_1, \ldots, X_n$ defined as

$$U = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} g(X_i, X_j). \tag{A.2}$$

Let $\|g\|_\infty \leq b$ for some $b > 0$. Then,

$$\mathbb{P}\left(|U - \mathbb{E}[U]| > t\right) \leq 2\exp\Big(-\frac{nt^2}{8b^2}\Big). \tag{A.3}$$

The proof follows directly from an implication of Azuma-Hoeffding's inequality. For example, see [46, Example 2.23] for a proof.

REFERENCES

[1] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[2] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Aug. 2010.

[3] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Statist.*, vol. 43, no. 1, pp. 177–214, 2015.

[4] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, Jun. 2012.

[5] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2009.

[6] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[7] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, pp. 3413–3430, Dec. 2011.

[8] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *Ann. Statist.*, vol. 39, no. 2, pp. 1069–1097, Apr. 2011.

[9] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, Mar. 2010.

[10] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," 2015, *arXiv:1509.03025*.

[11] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[12] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2973–2981.

[13] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, Jun. 2013, pp. 665–674.

[14] M. Hardt, "Understanding alternating minimization for matrix completion," in *Proc. IEEE 55th Annu. Symp. Found. Comput. Sci.*, Oct. 2014, pp. 651–660.

[15] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[16] D. Song, C. E. Lee, Y. Li, and D. Shah, "Blind regression: Nonparametric regression for latent variable models via collaborative filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2155–2163.

[17] Y. Li, D. Shah, D. Song, and C. L. Yu, "Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1760–1784, Mar. 2020.

[18] C. Borgs, J. Chayes, C. E. Lee, and D. Shah, "Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4715–4726.

[19] C. Borgs, J. T. Chayes, D. Shah, and C. L. Yu, "Iterative collaborative filtering for sparse matrix estimation," *Oper. Res.*, vol. 70, no. 6, pp. 3143–3175, Nov. 2022.

[20] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2012.

[21] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, 2011, Art. no. 025010.

[22] R. Tomioka, K. Hayashi, and H. Kashima, "Estimation of low-rank tensors via convex optimization," 2010, *arXiv:1010.0789*.

[23] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, "Statistical performance of convex tensor decomposition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 972–980.

[24] P. Jain and S. Oh, "Provable tensor factorization with missing data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–15.

[25] S. Bhojanapalli and S. Sanghavi, "A new sampling technique for tensors," 2015, *arXiv:1502.05023*.

[26] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Found. Comput. Math.*, vol. 16, no. 4, pp. 1031–1068, 2014.

[27] B. Barak and A. Moitra, "Noisy tensor completion via the sum-of-squares hierarchy," in *Proc. Conf. Learn. Theory*, 2016, pp. 417–445.

[28] A. Potechin and D. Steurer, "Exact tensor completion with sum-of-squares," in *Proc. Conf. Learn. Theory*, 2017, pp. 1619–1673.

[29] A. Montanari and N. Sun, "Spectral algorithms for tensor completion," *Commun. Pure Appl. Math.*, vol. 71, no. 11, pp. 2381–2425, 2018.

[30] D. Xia and M. Yuan, "On polynomial time methods for exact low-rank tensor completion," *Found. Comput. Math.*, vol. 19, no. 6, pp. 1–49, 2017.

[31] D. Xia, M. Yuan, and C.-H. Zhang, "Statistically optimal and computationally efficient low rank tensor completion from noisy entries," *Ann. Statist.*, vol. 49, no. 1, pp. 1–12, Feb. 2021.

[32] M. Yuan and C.-H. Zhang, "Incoherent tensor norms and their applications in higher order tensor completion," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6753–6766, Oct. 2017.

[33] S. Friedland and L.-H. Lim, "Nuclear norm of higher-order tensors," *Math. Comput.*, vol. 87, no. 311, pp. 1255–1281, 2018.

[34] E. Abbe, J. Fan, K. Wang, and Y. Zhong, "Entrywise eigenvector analysis of random matrices with low expected rank," *Ann. Statist.*, vol. 48, no. 3, pp. 1452–1474, Jun. 2020.

[35] Y. Chen, J. Fan, C. Ma, and K. Wang, "Spectral method and regularized MLE are both optimal for top-K ranking," *Ann. Statist.*, vol. 47, no. 4, p. 2204, 2019.

[36] Y. Zhong and N. Boumal, "Near-optimal bounds for phase synchronization," *J. Optim.*, vol. 28, no. 2, pp. 989–1016, Apr. 2018.

[37] C. Cai, G. Li, Y. Chi, H. V. Poor, and Y. Chen, "Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," *Ann. Statist.*, vol. 49, no. 2, pp. 944–967, Apr. 2021.

[38] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3345–3354.

[39] L. Ding and Y. Chen, "Leave-one-out approach for matrix completion: Primal and dual analysis," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 7274–7301, Nov. 2020.

[40] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–7.

[41] C. Cai, H. V. Poor, and Y. Chen, "Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1271–1282.

[42] D. Shah and C. L. Yu, "Iterative collaborative filtering for sparse noisy tensor estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 41–45.

[43] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[44] K. Balasubramanian, D. Gitelman, and H. Liu, "Nonparametric modeling of higher-order interactions via hypergraphons," *J. Mach. Learn. Res.*, vol. 22, pp. 1–146, Jan. 2021.

[45] N. Alon and J. H. Spencer, *The Probabilistic Method*. Hoboken, NJ, USA: Wiley, 2016.

[46] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019, vol. 48.

[47] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, Jan. 1963.

Devavrat Shah (Fellow, IEEE) is currently a Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology. His current research interests include interface of statistical inference and social data processing. His work has been recognized through the prize paper awards in machine learning, operations research, and computer science; and career prizes, including the 2010 Erlang Prize from the INFORMS Applied Probability Society and the 2008 ACM Sigmetrics Rising Star Award. He is a Distinguished Young Alumni of his alma mater IIT Bombay.

Christina Lee Yu (Member, IEEE) received the B.S. degree in computer science from the California Institute of Technology and the M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. She is currently an Assistant Professor with the School of Operations Research and Information Engineering, Cornell University. Prior to Cornell University, she was a Post-Doctoral Researcher at Microsoft Research New England. She received Honorable Mention for the 2018 INFORMS Dantzig Dissertation Award. She was a recipient of the 2021 Intel Rising Stars Award and the JPMorgan Faculty Research Award.