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

Devavrat Shah * Christina Lee Yu †

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 sparse tensor estimation 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 low 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.

1 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

*Massachusetts Institute of Technology, Cambridge, MA; devavrat@mit.edu
†Cornell University, Ithaca NY; cleeyu@cornell.edu
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 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.

1.1 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 [19, 20, 8], nuclear norm minimization [5, 6, 7, 27, 25, 23], gradient descent [19, 20, 9, 29, 13], alternating minimization [17, 15], and nearest neighbor style collaborative filtering [14, 28, 21, 3, 4]. These algorithms have been shown to be provably consistent as long as the number of observations is $\Omega(n \cdot r \cdot \text{poly}(\log n))$ for the noiseless setting where $r$ is the rank and $n$ is the number of rows and columns [19, 5]; similar results have been attained under additive Gaussian noise [20, 6] and generic bounded noise [8, 3]. Lower bounds show that $\Omega(n \cdot r \cdot \log(n))$ samples are necessary for consistent estimation, and $\Omega(n \cdot r)$ samples are necessary for exact recovery [6, 7], 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 [22, 12, 30, 31]. A $d$-order tensor where each dimension is length $n$ would be unfolded to a $n^{[d/2]} \times n^{[d/2]}$ matrix, resulting in a sample complexity of $O(n^{[d/2]} \cdot \text{poly}(\log n))$, 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 $O(n^{3} \cdot \text{poly}(\log n))$.

Subsequent progress has shown that it is feasible to improve upon naive sample complexity bound obtained by unfolding the tensor to a matrix when $d$ is odd [18, 2, 33, 1, 26, 24]. For a 3-order tensor, they provide consistent estimators requiring only a sample complexity of $\Omega(n^{3/2} \cdot \text{poly}(\log n))$. [18, 2] analyzes the alternating minimization algorithm for exact recovery of the tensor given noiseless observations and finite rank $r = \Theta(1)$. [1, 26] use the sum of squares (SOS) method, and [24]
introduces a spectral method. Both of these latter algorithms can handle noisy observations and overcomplete tensors where the rank is larger than the dimension. For a general $d$-order tensor these results translate into a sample complexity scaling as $O(n^{d/2})$, improving upon $O(n^{[d/2]}).$ [34] prove that tensor nuclear norm minimization can recover the underlying low-rank $d$-order tensor with $O(n^{d/2} \text{poly} \log n)$ samples in the noiseless setting; however, the algorithm is not efficiently computable as computing tensor nuclear norm is NP-hard [11].

[1] 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 $\Omega(d \text{rank})$ 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.

1.2 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 $\Omega(n^{3/2})$ 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 an iterative nearest neighbor collaborative filtering algorithm for sparse matrix estimation introduced in [3]. As discussed, the iterative collaborative filtering for matrix estimation of [3] applied to the $n \times n^2$ matrix obtained by unfolding the 3-order tensor does not work as it requires $\Omega(n^2)$ samples, far more than $\Omega(n^{3/2})$ required per best known computationally efficient estimator, for consistent estimation. 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 $\Omega(n^{3/2} + \kappa)$ samples for any positive $\kappa > 0$. And after computing these similarities, by using tensor structure through nearest neighbor, we can obtain consistent estimation.

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 5.4.

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 $\text{poly}(\varepsilon)$ and MEE converges to $\text{poly}(\varepsilon)$ 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 $\text{poly}(\varepsilon)$.

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 [3]. 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 [3]. 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 [3].

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 sample complexity of our algorithm for 3-order tensor is at most $pn^6$. As discussed in Section 5.4, with use of approximate nearest neighbors, these can be further improved and made truly implementable.

## 2 Preliminaries

Sparse tensor estimation 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 $\ell \in [d]$, $r_\ell$ is the dimension of the column space of $F(y)$. The Tucker rank is also the minimal values
Figure 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.

\[
\begin{align*}
F = (Q_1 \otimes \cdots \otimes Q_d) \cdot (\Lambda) &:= \sum_{k \in [r_1] \times [r_2] \times \cdots \times [r_d]} \Lambda(k) Q_1(\cdot, k_1) \otimes Q_2(\cdot, k_2) \cdots \otimes Q_d(\cdot, k_d),
\end{align*}
\] (2.1)

and depicted in Figure 2. The higher order SVD (HOSVD) specifies a unique Tucker decomposition in which the factor matrices \(Q_1 \ldots Q_d\) are orthonormal and correspond to the left singular vectors of the unfolded tensor along each mode [10].

If the CP-rank is \(r\), the Tucker-rank is bounded above by \((r, r, \ldots, r)\) by constructing a super-diagonal core tensor. If the Tucker rank is \((r_1, r_2, \ldots, r_t)\), the CP-rank is bounded by the number of nonzero entries in the core tensor, which is at most \(r_1 r_2 \cdots r_t / (\max \ell r_\ell)\) [10]. 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 \(\Lambda\) associated to the Tucker decomposition is superdiagonal, or equivalently enforcing that the latent factors in the minimal CP-decomposition are orthogonal. There does not always exist such an orthogonal CP-decomposition, however this class still includes all rank 1 tensors which encompasses the class of instances used to construct the hardness conjecture in [1]. Our results also extend beyond to general tensors as well, though the presentation is simpler in the orthogonal setting.
3 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 \( \sup_{u,v,w \in [0,1]} |f(\theta_u, \theta_v, \theta_w)| \leq 1 \).

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 \( (u,v,w) \). We shall assume that \( \sigma^2 = \mathbb{E}[(F(u,v,w) - \hat{F}(u,v,w))^2] \) is small, where \( \sigma^2 \) for an estimate \( \hat{F} \) is defined as

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

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

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

3.1 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 \( \ell_2 \) 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 \( \Lambda \) denote the diagonal \( r \times r \) matrix where \( \Lambda_{kk} = \lambda_k \). Let \( Q \) denote the \( r \times n \) matrix where \( Q_{kb} = q_k(\theta_{b_1}) q_k(\theta_{b_2}) \) 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 \( k \)th entry being 1 of appropriate dimension (here it is \( r \)).

\footnote{The notation of \( * \) is used to denote the missing observation. When convenient, we shall replace \( * \) by 0 for the purpose of computation.}
3.2 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 $f : [0,1]^3 \rightarrow \mathbb{R}$, it is said to have $\varepsilon$-approximate finite spectrum with rank $r$ for $\varepsilon \geq 0$ if there exists a symmetric function $f_r : [0,1]^3 \rightarrow \mathbb{R}$ such that

$$\sup_{\theta_u,\theta_v,\theta_w \in [0,1]} |f(\theta_u,\theta_v,\theta_w) - f_r(\theta_u,\theta_v,\theta_w)| \leq \varepsilon$$

where $r = \Theta(1)$ and $q_k(.)$ denotes the orthonormal $\ell_2$ 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 $\varepsilon$-approximate rank $r$ such that

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

(3.4)

3.3 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. This assumption can be relaxed as the main property that our algorithm and analysis use is the orthogonal decomposition of the unfolded tensors along each mode. In the general setting, 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.

3.4 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.

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 $\varepsilon$-approximate rank being $r = \Theta(1)$.

It is worth taking a note of the fact that to establish our result for the approximately finite spectrum setting, we perform a perturbation analysis wherein each observed entry is perturbed arbitrarily bounded by $\varepsilon$ 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.

7
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.

4 Algorithm

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

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 $\text{dist}(u, a)$ 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 $\sigma^2$ 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 $\theta_a, \theta_u$, and for randomly sampled $\theta_b, \theta_c$, 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,

$$
\mathbb{E}[(M(a, b, c) - M(u, b, c))^2 | \theta_a, \theta_u]
= \mathbb{E}[(F(a, b, c) - F(u, b, c))^2 | \theta_a, \theta_u] + \mathbb{E}[\epsilon_{abc}^2 + \epsilon_{ubc}^2]
$$

$$
= \mathbb{E}[(\sum_k \lambda_k(q_k(\theta_a) - q_k(\theta_u))q_k(\theta_b)q_k(\theta_c))^2 | \theta_a, \theta_u] + 2\sigma^2
$$

$$
= \mathbb{E}[(\sum_k \lambda_k^2(q_k(\theta_a) - q_k(\theta_u))^2q_k(\theta_b)q_k(\theta_c))^2 | \theta_a, \theta_u] + 2\sigma^2
$$

$$
= \sum_k \lambda_k^2(q_k(\theta_a) - q_k(\theta_u))^2 + \sigma^2
$$

$$
= \|\Lambda Q(e_a - e_u\|_2^2 + 2\sigma^2,
$$

where we use the fact that $q_k(\cdot)$ are orthonormal. This suggests that approximating $\text{dist}(u, a)$ 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 $n \times (n/2)$ “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 $b \leq c$. 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; and allow for noise terms in each entry to have different distributions and variance. We do so by constructing a graph using the non-zero pattern of the matrix as an adjacency matrix. This mirrors the idea from [3] for matrix estimation, which approximates distances by comparing expanded local neighborhoods in the graph representing the sparsity pattern of the unfolded or flattened tensor. A key distinction is
that the tensor setting requires an important modification of how one constructs the local breadth-first-search (BFS) trees, described in step 3 below.

4.1 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, \( \Omega_1 \) and \( \Omega_2 \). Tensors \( M_1 \) and \( M_2 \) contain information from the subset of the data in \( M \) associated to \( \Omega_1 \) and \( \Omega_2 \) 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 \( V_A \) denote the set of coordinate pairs within set 1 consisting of distinct coordinates, i.e. \( V_A = \{(b, c) \in [n/2]^2 \text{ s.t. } b < c\} \). Let \( V_B \) denote the set of coordinate pairs within set 2 consisting of distinct coordinates, i.e. \( V_B = \{(b, c) \in ([n] \setminus [n/2]^2) \text{ s.t. } b < c\} \). The sizes of |\( V_A \)| and |\( V_B \)| are both equal to \((n/2)^2\). We define \( M_A \) to be the \( n \)-by-\((n/2)^2\) matrix taking values \( M_A(a, (b, c)) = M_1(a, b, c) \), where each row corresponds to an original coordinate of the tensor, and each column corresponds to a pair of coordinates \((b, c) \in V_A\) from the original tensor. We define \( M_B \) to be the \( n \)-by-\((n/2)^2\) 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 \((b, c) \in V_B\) 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.

Step 2: Construct Bipartite Graph from \( \Omega_1, M_A \). We define a bipartite graph corresponding to the flattened matrix \( M_A \). Construct a graph with vertex set \([n] \cup V_A\). There is an edge between vertex \( a \in [n] \) and vertex \((b, c) \in V_A\) 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.

Step 3: Expanding the Neighborhood. Consider the graph constructed from \( \Omega_1, M_A \). 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 \(U_{u,s} \subseteq \mathcal{V}_A\) to denote the set of vertices at depth \((2s-1)\) of the BFS tree, and we use \(S_{u,s} \subseteq [n]\) to denote the set of vertices at depth \(2s\) of the BFS tree. Let \(B_{u,s} \subset [n] \cup \mathcal{V}_A\) denote the set of vertices which are visited in the first \(s\) layers of the BFS tree,

\[ B_{u,s} = \bigcup_{h \in [s/2]} S_{u,h} \cup \bigcup_{l \in [s/2]} U_{u,l}. \]

We will overload notation and sometimes use \(B_{u,s}\) 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.

\[ B_{u,s} = \bigcup_{h \in [s/2]} S_{u,h} \cup \{ x \in [n] \text{ s.t. } \exists (y, z) \in \bigcup_{l \in [s/2]} U_{u,l} \text{ satisfying } x \in \{y, z\}\}. \]

Let \(G(B_{u,s})\) 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 \(B_{u,s}\), the latent variables \(\{\theta_a\}_{a \in B_{u,s}}\) and the edge weights \(\{M_1(a, b, c)\}_{a, b, c} \in B_{u,s}\).

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 \(S_{u,s}\), where the \(a\)-th coordinate is equal to the product of weights along the path from \(u\) to \(a\) in the BFS tree for \(a \in S_{u,s}\). Similarly, let \(W_{u,s} \in [0, 1]^{\mathcal{V}_A}\) be associated to set \(U_{u,s}\), 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 \((b, c) \in U_{u,s}\). 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,

\[ N_{u,s}(a) = M_A(a, \pi_u(a))W_{u,s}(\pi_u(a))I_{a \in S_{u,s}} \]
\[ W_{u,s}(b, c) = M_A(\pi_u(b, c), (b, c))N_{u,s-1}(\pi_u(b, c))I_{(b, c) \in U_{u,s}} \]

and \(N_{u,0} = e_u\). Let \(\tilde{N}_{u,s}\) denote the normalized vector \(\tilde{N}_{u,s} = N_{u,s}/|S_{u,s}|\) and let \(\tilde{W}_{u,s}\) denote the normalized vector \(\tilde{W}_{u,s} = W_{u,s}/|U_{u,s}|\). Figure 3(d) illustrates the neighborhood sets and vectors for a valid BFS tree.

**Step 4: Computing the distances using** \(M_B\). Let

\[ t = \left\lfloor \frac{\ln(n)}{2 \ln(p^2 n^3)} \right\rfloor. \tag{4.1} \]

A heuristic for the distance would be

\[ \text{dist}(u, v) \approx \frac{1}{|V_B|^2} (\tilde{N}_{u,t} - \tilde{N}_{v,t})M_BM_B^T (\tilde{N}_{u,t} - \tilde{N}_{v,t}) \tag{4.2} \]
\[ = \frac{1}{|V_B|^2} \sum_{(a, \beta) \in V_B} \sum_{a, b \in [n]^2} (\tilde{N}_{u,t}(a) - \tilde{N}_{v,t}(a))M_B(a, \alpha, \beta)M_B(b, \alpha, \beta)(\tilde{N}_{u,t}(b) - \tilde{N}_{v,t}(b)). \]

For technical reasons that facilitate cleaner analysis, we use the following distance calculations. There are two deviations from the equation in (4.2). First we exclude \(a = b\) from the summation. Second we exclude coordinates for \(\alpha\) or \(\beta\) that have been visited previously in \(B_{u,2t}\) or \(B_{v,2t}\). Define
Figure 3: Consider a symmetric 3-order tensor with \( n = 6 \), and the observation set \( \Omega_1 = \{(1, 3, 4), (1, 2, 5), (2, 3, 4), (2, 4, 5), (2, 5, 6), (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 \((2, 5, 4)\) and \((3, 4, 2)\) are not valid, as coordinates 2 and 4 have both been visited in layer 2 by the vertices \((2, 5)\) and \((3, 4)\). For the BFS tree rooted at vertex 4, edge \((2, 5, 1)\) is not valid as coordinate 1 has been visited in layer 2 by the vertex \((1, 3)\) and edge \((6, 3, 5)\) is not valid as coordinates 4 and 5 have both been visited in layer 2.
Figure (d) depicts the sets \( S_{u,s} \) and \( U_{u,s} \) along with the neighborhood vectors \( N_{u,s} \) and \( W_{u,s} \) for a specific valid BFS tree rooted at vertex \( u = 4 \).
distance as
\[
\text{dist}(u, v) = (Z_{uu} + Z_{vv} - Z_{uv} - Z_{vu}),
\]
\[
Z_{uv} = \frac{1}{|\mathcal{V}_B(u, v, t)|^2 |S_{u,t}| |S_{v,t}|} \sum_{(\alpha, \beta) \in \mathcal{V}_B(u, v, t)} \bar{T}_{uv}(\alpha, \beta),
\]
\[
\mathcal{V}_B(u, v, t) = \{(\alpha, \beta) \in \mathcal{V}_B \text{ s.t. } \alpha \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t} \text{ and } \beta \notin \mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t}\},
\]
\[
T_{uv}(\alpha, \beta) = \sum_{a \neq b \in [n]} N_{u,t}(a)N_{v,t}(b)M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta)).
\]

**Step 5: Averaging datapoints to produce final estimate.** Let \(\Omega_{uvw}\) 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 \(\text{dist}(u, a), \text{dist}(v, b), \text{dist}(w, c)\) are all less than some chosen threshold parameter \(\eta\). The final estimate averages the datapoints corresponding to indices in \(\Omega_{uvw}\),
\[
\hat{F}(u, v, w) = \frac{1}{|\Omega_{uvw}|} \sum_{(a, b, c) \in \Omega_{uvw}} M_2(a, b, c).
\]

### 4.2 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. As a result 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.

### 5 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 by 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.

#### 5.1 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 (4.1). 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,
\[
\text{MSE} = O(n^{-(\kappa - \psi)}) = O\left(\frac{n^\psi}{(p^2 n^3)^{1/2}}\right),
\]
and
\[
\|F - \hat{F}\|_{\text{max}} = O(n^{-(\kappa - \psi)/2}),
\]
with probability \( 1 - O\left(n^4 \exp(-\Theta(n^{2\psi}))\right) \).

5.2 Approximately finite spectrum

For approximate rank \( r \) model, we establish a natural perturbation result for the algorithm. Specifically, if the underlying model has \( \varepsilon \)-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 \( \text{poly}(\varepsilon) \).

Theorem 5.2. We assume that the function \( f \) has \( \varepsilon \)-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 (4.1), it follows that \( t = \lceil \frac{1}{4\varepsilon} \rceil \). 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,
\[
\text{MSE} = O(n^{-(\kappa - \psi)} + t\varepsilon(1 + \varepsilon)^{2t-1} + t^2 \varepsilon^2 (1 + \varepsilon)^{4t-2})
\]
\[
= O\left(\frac{n^\psi}{(p^2 n^3)^{1/2}} + t\varepsilon(1 + \varepsilon)^{2t-1} + t^2 \varepsilon^2 (1 + \varepsilon)^{4t-2}\right),
\]
and
\[
\|F - \hat{F}\|_{\text{max}} = O(n^{-\kappa - \psi}/2 + t\varepsilon(1 + \varepsilon)^{2t-1} + \sqrt{t\varepsilon(1 + \varepsilon)^{2t-1}}),
\]
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 \( \|F\|_{\text{max}} \leq 1 \), the bound is meaningful when \( \varepsilon < 1 \), in which case the dominating term of the additional error due to the perturbation is linear in \( \varepsilon \), 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 \( \varepsilon \)-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 \( \varepsilon \). 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 \( \varepsilon \), the estimation error is perturbed by at most \( \text{poly}(\varepsilon) \). As a byproduct, our result proves that our estimator that is robust to arbitrary deterministic bounded noise in the observations.
5.3 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 $(a, \alpha, \beta)$ and $(b, \alpha, \beta)$ 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 [3], 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 (4.1). For any arbitrarily small $\psi \in (0,\min(\kappa,\frac{1}{2}))$, 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

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

and

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

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

5.4 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 iterative collaborative filtering for matrix estimation [3].

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 \( \theta \). 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 [28]. Similar extensions as presented in [28] 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 \( \theta \) 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, we believe that our algorithm and analysis can be extended to general \( d \)-order tensors. A naive calculation seems to imply that \( \omega(n^{d/2}) \) is the connectivity threshold for the graph corresponding to a flattening of a \( d \)-order tensor.

6 Proof

In this section, we present the proof for Theorem 5.1. The proof outline is similar to the matrix setting in [3], in that the core of the analysis is proving that the distance function as defined in (4.3) 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 6.2 requires a new argument beyond the simple martingale argument in the matrix setting. This involves a careful application of concentration of U-statistics. Furthermore, the concentration of the distance calculation in Eq (6) as analyzed in section 6.4 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.

6.1 Analyzing Noisy Nearest Neighbors

We start by stating an important Lemma 6.1, adapted from [3] 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 (4.5), 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 $\eta$ allows for tradeoff between bias and variance of the algorithm.

We first argue that the data-driven distance estimates $\text{dist}$ will concentrate around an ideal data-independent distance $d(\theta_u, \theta_v)$ for $d : [0,1]^2 \rightarrow \mathbb{R}_+$. We subsequently argue that the nearest neighbor estimate produced by (4.5) using $d(\theta_u, \theta_v)$ in place of $\text{dist}(u,v)$ will yield a good estimate by properly choosing the threshold $\eta$ 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 $d(\theta_u, \theta_a) \leq \eta$, $d(\theta_v, \theta_b) \leq \eta$ and $d(\theta_w, \theta_c) \leq \eta$ needs to be sufficiently large. We formalize the above stated desired properties.

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

Property 6.2 (Good Distance Estimation). For some $\Delta > 0$, we call distance $\hat{d} : [n]^2 \rightarrow \mathbb{R}_+$ a $\Delta$-good estimate for ideal distance $d : [0,1]^2 \rightarrow \mathbb{R}_+$, if $|d(\theta_u, \theta_a) - \hat{d}(u,a)| \leq \Delta$ 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 meas : $\mathbb{R}_+ \rightarrow \mathbb{R}_+$ if for any $u \in [n]$ and $\eta' > 0$, $\frac{1}{n} \sum_{a \in [n]} I(d(u,a) \leq \eta') \geq \text{meas}(\eta')$.

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 $\eta, \Delta$, and $\eta' = \eta - \Delta$; in particular $\hat{d}$ is a bias-good distance function, $\hat{d} = \text{dist}$ as estimated from $M_2$ is a $\Delta$-good distance estimate for $d$, and $\{\theta_u\}_{u \in [n]}$ is meas-represented. Then noisy nearest neighbor estimate $\hat{F}$ computed according to (4.5) satisfies

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

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

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

with probability at least

$$1 - \exp\left(-\frac{1}{2}\delta^2p \text{meas}(\eta - \Delta)n^3\right) - \exp\left(-\delta^2(1 - \delta)p \text{meas}(\eta - \Delta)n^3\right) - 3n\alpha_1 - \alpha_2.$$
The proof of Lemma 6.1 is a modification from [3] and is included in the Appendix.

### 6.2 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 \( \tilde{d} = \text{dist} \) computed according to (4.3) with \( t \) defined in (4.1). 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. (3.4).

**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

\[
d(\theta_u, \theta_v) = \| \Lambda^{t+1} Q(e_u - e_v) \|_2^2 = \sum_{k=1}^{r} \lambda_k^{2(t+1)} (q_k(\theta_u) - q_k(\theta_v))^2 .
\] (6.1)

Recall that \( t \) is defined in (4.1). Since \( p = n^{-3/2+\kappa} \) and \( \kappa \in (0, \frac{1}{2}) \), we have that

\[
t = \left\lceil \frac{\ln(n)}{2 \ln(p^2 n^\gamma)} \right\rceil = \left\lceil \frac{1}{4\kappa} \right\rceil .
\] (6.2)

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

\[
|f(\theta_u, \theta_v, \theta_w) - f(\theta_a, \theta_b, \theta_c)| \leq |f(\theta_u, \theta_v, \theta_w) - f(\theta_a, \theta_v, \theta_w)| + |f(\theta_a, \theta_v, \theta_w) - f(\theta_a, \theta_b, \theta_w)| + |f(\theta_a, \theta_b, \theta_w) - f(\theta_a, \theta_b, \theta_c)| .
\] (6.3)

Now

\[
|f(\theta_u, \theta_v, \theta_w) - f(\theta_a, \theta_v, \theta_w)| = |\sum_k \lambda_k(q_k(\theta_u) - q_k(\theta_a)) q_k(\theta_v) q_k(\theta_w)| \leq B^2 |\sum_k \lambda_k(q_k(\theta_u) - q_k(\theta_a))|
\]

\[
= B^2 \| \Lambda Q(e_u - e_a) \|_1
\]

\[
\leq B^2 \sqrt{r} |\lambda_r|^{-t} \| \Lambda^{t+1} Q(e_u - e_a) \|_2
\]

\[
= B^2 \sqrt{r} |\lambda_r|^{-t} \sqrt{d(\theta_u, \theta_a)} .
\] (6.4)

In above, (a) follows from the \( \| q_k(\cdot) \|_\infty \leq B \) for all \( k \). Repeating this argument to bound the other terms in (6.3), we obtain that

\[
|f(\theta_u, \theta_v, \theta_w) - f(\theta_a, \theta_b, \theta_c)| \leq 3B^2 \sqrt{r} |\lambda_r|^{-t} \max(\sqrt{d(\theta_u, \theta_a)}, \sqrt{d(\theta_v, \theta_b)}, \sqrt{d(\theta_w, \theta_c)}) \leq 3B^2 |\lambda_r|^{-t} \sqrt{r \eta} \equiv \text{bias}(\eta) .
\] (6.5)

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

17
Good distance estimate \( \hat{d} \) and Property 6.2. We state the following Lemma whose proof is delegated to Section 7.

**Lemma 6.2.** Given \( f \) with rank \( r \), assume that \( p = n^{-3/2 + \kappa} \) for \( \kappa \in (0, \frac{1}{2}) \). Let \( \hat{d} = \text{dist} \) as defined in (4.3). Then for any \((u, a) \in [n]^2\), for any \( \psi \in (0, \kappa) \),
\[
|d(\theta_u, \theta_a) - \hat{d}(u, a)| = O\left(n^{-2(p)}\right),
\]
with probability at least \(1 - O\left(\exp(-n^{2p}(1-o(1)))\right)\).

Lemma 6.2 implies that property 6.2 holds with probability \(1 - o(1)\) for \( \Delta = \Theta\left(n^{-2(p)}\right) \) when \( f \) has rank \( r \).

**Lemma 6.3.** Given \( f \) with \( \varepsilon \)-approximate rank \( r \) for \( \varepsilon \geq 0 \), assume that \( p = n^{-3/2 + \kappa} \) for \( \kappa \in (0, \frac{1}{2}) \). Let \( \hat{d} = \text{dist} \) as defined in (4.3). Then for any \((u, a) \in [n]^2\), for any \( \psi \in (0, \kappa) \),
\[
|d(\theta_u, \theta_a) - \hat{d}(u, a)| = O\left(n^{-2(p)}\right) + O\left(t\varepsilon(1 + \varepsilon)^{2r-1} + t^2\varepsilon^2(1 + \varepsilon)^{4r-2}\right),
\]
with probability at least \(1 - O\left(\exp(-n^{2p}(1-o(1)))\right) - O\left(n^{-6}\right)\).

Lemma 6.3 implies that property 6.2 holds with probability \(1 - o(1)\) for \( \Delta = \Theta\left(n^{-2(p)} + t\varepsilon(1 + \varepsilon)^{2r-1} + t^2\varepsilon^2(1 + \varepsilon)^{4r-2}\right) \), when \( f \) has \( \varepsilon \)-approximate rank \( r \).

**Sufficient representation and Property 6.3.** Since \( f \) is \( L \)-Lipschitz, the distance \( d \) as defined in (6.1) is bounded above by the squared \( \ell_2 \) distance:
\[
d(\theta_u, \theta_v) = \|A^{t+1}Q(e_u - e_v)\|_2^2
\leq |\lambda|^2\|AQ(e_u - e_v)\|_2^2
\]
\[
= |\lambda|^2\left(\sum_{k=1}^r \lambda_k^2(q_k(\theta_u) - q_k(\theta_v))^2\right)
\]
\[
= |\lambda|^2\left(\sum_{k=1}^r \lambda_k^2(q_k(\theta_u) - q_k(\theta_v))^2\left(\int_0^1 q_k(\theta_a)^2d\theta_a\right)\left(\int_0^1 q_k(\theta_b)^2d\theta_b\right)\right)
\]
\[
= |\lambda|^2\left(\sum_{k=1}^r \lambda_k^2\int_0^1 \int_0^1 (q_k(\theta_u)q_k(\theta_a)q_k(\theta_b) - q_k(\theta_v)q_k(\theta_a)q_k(\theta_b))^2d\theta_ad\theta_b\right)
\]
\[
\overset{(a)}{=} |\lambda|^2\left(\int_0^1 \int_0^1 (f(\theta_u, \theta_a, \theta_b) - f(\theta_v, \theta_a, \theta_b))^2d\theta_ad\theta_b\right)
\]
\[
\leq |\lambda|^2L^2|\theta_u - \theta_v|^2,
\]
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 \( \theta_u \) are sampled i.i.d. uniformly over \([0, 1]\). Therefore, for any \( \theta_u \in [0, 1] \), for any \( v \in [n] \) and \( \eta' > 0 \),
\[
\mathbb{P}\left(d(\theta_u, \theta_v) \leq \eta' \mid \theta_u\right) \geq \mathbb{P}\left(|\lambda|^2L^2|\theta_u - \theta_v|^2 \leq \eta' \mid \theta_u\right)
\]
\[
= \mathbb{P}\left(|\theta_u - \theta_v| \leq \frac{\sqrt{\eta'}}{|\lambda|^2L} \mid \theta_u\right)
\]
\[
\geq \min\left(1, \frac{\sqrt{\eta'}}{|\lambda|^2L}\right).
\]
Let us define

$$\text{meas}(\eta') = \frac{(1 - \delta)\sqrt{\eta'}}{|\lambda_1|^tL}$$

(6.10)

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

$$\mathbb{P}\left(\frac{1}{n-1} \sum_{u \in [n]\setminus u} \mathbb{I}(d(u, a) \leq \eta') \leq \text{meas}(\eta') \mid \theta_u\right) \leq \exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^tL}\right).$$

By using union bound over all $n$ indices, it follows that for any $\eta' \in (0, |\lambda_1|^{2t}L^2)$, with probability at least $1 - n \exp\left(-\frac{\delta^2(n-1)\sqrt{\eta'}}{2|\lambda_1|^tL}\right)$, property 6.3 is satisfied with meas as defined in (6.10).

**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 $\alpha_1 = O\left(\exp\left(-n^{2\psi}(1-o(1))\right)\right)$ 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 $\alpha_2 = n \exp\left(-\frac{\delta^2(n-1)\sqrt{\eta' - \Delta}}{2|\lambda_1|^tL}\right)$ with distance estimate $d = \text{dist}$ defined in (4.3) with

\[
\begin{align*}
\mathbf{d}(\theta_u, \theta_v) &= ||\mathbf{A}^{1+1}Q(e_u - e_v)||_2^2, \\
\mathbf{bias} &= 3B^2|\lambda_\tau|^{-l} \sqrt{\eta'}, \\
\Delta &= \Theta\left(n^{-(\kappa - \psi)}\right), \\
\text{meas}(\eta') &= \frac{(1 - \delta)\sqrt{\eta'}}{|\lambda_1|^tL},
\end{align*}
\]

(6.11)

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

\[
\text{MSE}(\hat{F}) \leq 9B^4|\lambda_\tau|^{-2t}r(\eta + \Delta) + \frac{\sigma^2L^3|\lambda_1|^{3t}}{(1 - \delta)^4p\left(\sqrt{\eta - \Delta}n\right)^3} \\
+ \exp\left(-\frac{\delta^2(1 - \delta)^3p\left(\sqrt{\eta - \Delta}n\right)^3}{2L^3|\lambda_1|^{3t}}\right) \\
+ nO\left(\exp(-n^{2\psi}(1-o(1)))\right) + n \exp\left(-\frac{\delta^2(n-1)(\sqrt{\eta - \Delta})}{2|\lambda_1|^tL}\right).
\]

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

$$|\hat{F}(u, v, w) - f(\theta_u, \theta_v, \theta_w)| \leq 3B^2|\lambda_\tau|^{-l} \sqrt{r(\eta + \Delta)} + \delta'$$

(6.12)

with probability at least

\[
\begin{align*}
1 - \exp\left(-\frac{\delta^2(1 - \delta)^3p\left(\sqrt{\eta - \Delta}n\right)^3}{2L^3|\lambda_1|^{3t}}\right) - \exp\left(-\frac{\delta^2(1 - \delta)^3p\left(\sqrt{\eta - \Delta}n\right)^3}{L^3|\lambda_1|^{3t}}\right) \\
- nO\left(\exp(-n^{2\psi}(1-o(1)))\right) - n \exp\left(-\frac{\delta^2(n-1)(\sqrt{\eta - \Delta})}{2|\lambda_1|^tL}\right).
\]

19
By selecting \( \eta = \Theta(\Delta) = \Theta(n^{-(\kappa-\psi)}) \) with a large enough constant so that \( \eta - \Delta = \Theta(\eta) \), it follows that by the conditions that \( \psi > 0 \) and \( \kappa < \frac{1}{2} \),

\[
\begin{align*}
\eta \pm \Delta &= \Theta(n^{-(\kappa-\psi)}), \\
p(\sqrt{\eta - \Delta} n) &= \Theta(n^{\frac{3}{2} - \frac{1}{2} + \frac{3\psi}{2}}) = \Omega(n^{\frac{5}{2}}), \\
n(\sqrt{\eta - \Delta} n) &= \Theta(n^{1 - \frac{3\psi}{2}}) = \Omega(n^{\frac{3}{2}}).
\end{align*}
\]

By substituting this choice of \( \eta \) and \( \delta = \frac{1}{2} \), it follows that

\[
\text{MSE}(\hat{F}) = O\left(n^{-(\kappa-\psi)}\right).
\]

By choosing \( \delta' = n^{-(\kappa-\psi)/2} \) such that \( \delta' = \Theta(\sqrt{\eta}) \) and \( \delta'^2 p(\sqrt{\eta - \Delta} n) = \Omega(n^{\frac{3}{2}}) = \Omega(n^{2\psi}) \) because \( \psi < \frac{3}{8} \). Therefore, by substituting into (6.12), it follows that for any given \( (u, v, w) \in [n]^3 \), with probability \( 1 - O(n\exp(-\Theta(n^{2\psi}))) \),

\[
|\hat{F}(u, v, w) - f(\theta_u, \theta_v, \theta_w)| = O\left(n^{-(\kappa-\psi)/2}\right).
\]

Using union bound over choices of \( (u, v, w) \in [n]^3 \), it follows that the maximum entry-wise error is bounded above by \( O\left(n^{-(\kappa-\psi)/2}\right) \) with probability \( 1 - O(n^4\exp(-\Theta(n^{2\psi}))) \). 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 \( \alpha_1 = O\left(\exp(-n^{2\psi}(1 - o(1)) + n^{-6})\right) \) 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 \( \alpha_2 = n\exp\left(-\frac{\delta^2(n-1)\sqrt{\eta - \Delta}}{2L|\lambda_1|^2}\right) \) with distance estimate \( \hat{d} = \text{dist} \) defined in (4.3) with

\[
\begin{align*}
d(\theta_u, \theta_v) &= \|\Lambda^{t+1}Q(e_u - e_v)\|_2^2, \\
\text{bias}(\eta) &= 3B^2|\lambda_1|^{-t}\sqrt{\eta}, \\
\Delta &= \Theta(n^{-(\kappa-\psi)} + t\varepsilon(1 + \varepsilon)^{2t-1} + t\varepsilon^2(1 + \varepsilon)^{4t-2}), \\
\text{meas}(\eta') &= \frac{(1 - \delta)\sqrt{\eta}}{|\lambda_1|^tL},
\end{align*}
\]

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

\[
|\hat{F}(u, w, v) - f_r(\theta_u, \theta_v, \theta_w)| \leq 3B^2|\lambda_1|^{-t}\sqrt{r(\eta + \Delta)} + \delta',
\]

with probability at least

\[
1 - \exp\left(-\frac{\delta^2(1 - \delta)^3p\left(\sqrt{\eta - \Delta} n\right)^3}{2L^3|\lambda_1|^{3t}}\right) - \exp\left(-\frac{\delta^2(1 - \delta)^4p\left(\sqrt{\eta - \Delta} n\right)^3}{L^3|\lambda_1|^{3t}}\right) - nO\left(\exp(-n^{2\psi}(1 - o(1))) + n^{-6}\right) - n\exp\left(-\frac{\delta^2(n - 1)\sqrt{\eta - \Delta}}{2|\lambda_1|^tL}\right).
\]

20
By selecting
\[ \eta = \Delta + \min \left( \Delta, |\lambda_1|^{2t}L^2 \right), \] (6.17)
it follows by the conditions \( \psi > 0 \) and \( \kappa < \frac{1}{2} \) that
\[
\begin{align*}
\eta + \Delta &= \Theta\left(n^{-\left(\kappa-\psi\right)} + t\varepsilon(1 + \varepsilon)^{2t-1} + t^2\varepsilon^2(1 + \varepsilon)^{4t-2}\right), \\
\eta - \Delta &= \Omega\left(n^{-\left(\kappa-\psi\right)}\right), \\
p(\sqrt{\eta - \Delta}n)^3 &= \Omega(n^{\frac{5}{2}}), \\
n\sqrt{\eta - \Delta} &= \Omega(n^{\frac{3}{2}}).
\end{align*}
\]

By choosing \( \delta' = n^{-\left(\kappa-\psi\right)/2} \) such that \( \delta' = O(\sqrt{\eta}) \) and \( \delta'^2 p(\sqrt{\eta - \Delta}n)^3 = \Omega(n^{\frac{3}{2}}) = \Omega(n^{2\psi}) \) because \( \psi < \frac{3}{8} \). Therefore, by substituting into (6.16), it follows that for any given \( \langle u, v, w \rangle \in [n]^3 \), with probability \( 1 - O(n \exp(-\Theta(n^{2\psi}))) - O(n^{-5}) \),
\[
|\hat{F}(u, v, w) - f(\theta_u, \theta_v, \theta_w)| \leq |\hat{F}(u, v, w) - f_r(\theta_u, \theta_v, \theta_w)| + |f_r(\theta_u, \theta_v, \theta_w) - f(\theta_u, \theta_v, \theta_w)| \leq O\left(n^{-\left(\kappa-\psi\right)/2} + t\varepsilon(1 + \varepsilon)^{2t-1} + \sqrt{t\varepsilon(1 + \varepsilon)^{2t-1}}\right),
\] (6.18)
\[
\text{where the bias between } f_r \text{ and } f \text{ is bounded by } \varepsilon, \text{ and dominated by the bound between } \hat{F} \text{ and } F_r. \]

Using union bound over choices of \( \langle u, v, w \rangle \in [n]^3 \), it follows that the maximum entry-wise error is bounded above by \( O\left(n^{-\left(\kappa-\psi\right)/2}\right) \) with probability \( 1 - O(n \exp(-\Theta(n^{2\psi}))) - O(n^{-2}) \).

The bound on MSE also follows by substituting \( \delta = \frac{1}{2} \) and the same choice of \( \eta \) from (6.17) 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
\[
\text{MSE}(\hat{F}) = O\left(n^{-\left(\kappa-\psi\right)/2} + t\varepsilon(1 + \varepsilon)^{2t-1} + \sqrt{t\varepsilon(1 + \varepsilon)^{2t-1}}\right). \] (6.20)

This completes the proof of Theorem 5.2.

### 6.3 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 \( \langle a, b \rangle \in \mathcal{Y}^2 \), the estimate \( \hat{F}(a, b) \) follows the same exact computation as described in Section 4.1. As a result it follows from Theorem 5.1 that with high probability,
\[
\max_{\langle a, b, c \rangle \in \mathcal{Y}^3} |F(a, b, c) - \hat{F}(a, b, c)| = O\left(n^{-\left(\kappa-\psi\right)/2}\right).
\]

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

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

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

\[
|\hat{F}(u, v, w) - F(u, v, w)| \leq |\hat{F}(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w))| + |F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(u, v, w)|.
\]

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,

\[
|F(\hat{\zeta}(u), \hat{\zeta}(v), \hat{\zeta}(w)) - F(u, v, w)| \leq 3B^2\sqrt{\lambda_1}r^{-1}\max\left\{d(\theta_u, \theta_{\hat{\zeta}(u)}), d(\theta_v, \theta_{\hat{\zeta}(u)}), d(\theta_w, \theta_{\hat{\zeta}(w)})\right\}.
\]

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 \( \alpha_1 = O\left(\exp(-n^2\psi(1 - o(1)))\right) \) 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

\[
d(\theta_u, \theta_{\hat{\zeta}(u)}) \overset{(a)}{\leq} \hat{d}(u, \hat{\zeta}(u)) + \Delta \\
\overset{(b)}{\leq} \hat{d}(u, \hat{\zeta}(u)) + \Delta \\
\overset{(c)}{\leq} d(\theta_u, \theta_{\hat{\zeta}(u)}) + 2\Delta \\
\overset{(d)}{=} \min_{a \in A} d(\theta_u, \theta_a) + 2\Delta,
\]

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 \( \zeta \).

To bound \( \min_{a \in A} d(\theta_u, \theta_a) \), we use (6.9) from property 6.3 to show that for any \( u \in [n] \), \( \eta = \Theta(n^{-\kappa + \psi}) \), and \( y = \Omega((p^2n^3)^{1/4}) = \Omega(n^{\kappa/2}) \)

\[
\mathbb{P}\left(\min_{a \in \mathcal{Y}} d(\theta_u, \theta_a) > \eta \mid \theta_u\right) = \prod_{a \in \mathcal{Y}} \mathbb{P}\left(d(\theta_u, \theta_a) > \eta \mid \theta_a\right)
\leq \left(1 - \frac{\eta}{|\lambda_1|^y L}\right)^y
\leq \exp(-\frac{y\eta}{|\lambda_1|^L}) = \exp(-\Theta(n^{\psi/2})).
\]

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 \( O(n^{-(\kappa - \psi)}) \). 

\[\square\]
7 Proving distance estimate is close

In this section we argue that the distance estimate as defined in (4.3) is close to an ideal distance as claimed in the Lemma 6.2.

7.1 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 \( s \geq 1 \), \( \mathcal{U}_{u,s} \subseteq \mathcal{V}_A \) 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, \( \mathcal{B}_{u,s} = \cup_{t \in [s/2]} \mathcal{U}_{u,t} \cup_{t \in [s/2]} \mathcal{S}_{u,t} \cup \mathcal{G}(\mathcal{B}_{u,s}) \) 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 \( \{ M_I(a,b,c) \}_{a,b,c \in \mathcal{B}_{u,s}} \). 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 \( W_{u,s} = \mathcal{S}_{u,s} / |\mathcal{U}_{u,s}| \) for \( u \in [n] \), \( s \geq 1 \).

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,

\[
N_{u,s}(a) = M_A(a, \pi_u(a)) W_{u,s}(\pi_u(a)) I_{a \in \mathcal{S}_{u,s}} \\
W_{u,s}(b,c) = M_A(\pi_u(b,c), (b,c)) N_{u,s-1}(\pi_u(b,c)) I_{(b,c) \in \mathcal{U}_{u,s}}
\]

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 (4.1). For a given \( \delta \in (0, \frac{1}{2}) \) and for any \( u \in [n] \), the following holds with probability \( 1 - O \left( n \exp \left( - \Theta(n^{2\kappa}) \right) \right) \):

- For all \( s \in [t-1] \),
  \[
  |\mathcal{S}_{u,s}| \leq \left[ (1-\delta)^{2s-2-3s} n^{2\kappa s} (1-o(1)), (1+\delta)^{2s-2-s} n^{2\kappa s} \right], \tag{7.1}
  \]
  and for \( s = t \),
  \[
  |\mathcal{S}_{u,t}| \leq \left[ (1-\delta)^{2t-2-3t-1} n^{2\kappa t} (1-o(1)), (1+\delta)^{2t-2-t} n^{2\kappa t} \right]. \tag{7.2}
  \]
- For \( s \in [t] \),
  \[
  |\mathcal{U}_{u,s}| \leq \left[ (1-\delta)^{2s-1-2-3s} n^{\frac{1}{2}+\kappa(2s-1)} (1-o(1)), (1+\delta)^{2s-1-2-s} n^{\frac{1}{2}+\kappa(2s-1)} \right]. \tag{7.3}
  \]
The set of single coordinate vertices visited within depth 2t is o(n),

\[ |\bigcup_{\ell=0}^{t} S_{u,\ell}| = o(n). \]  

(7.4)

**Proof.** First observe that if \( t \) is as defined in (4.1) with \( \kappa \in (0, \frac{1}{2}) \), then

\[ t = \left\lfloor \frac{\ln(n)}{2 \ln(p^2 n^3)} \right\rfloor = \left\lfloor \frac{1}{4\kappa} \right\rfloor \]

such that

\[ \frac{1}{4\kappa} \leq t < \frac{1}{4\kappa} + 1. \]  

(7.5)

Note that \( t \) is constant with respect to \( n \).

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

\[ \Pr \left( |S_{u,s}| \notin (1 \pm \delta)([n] \setminus B_{u,2s-1})(1 - (1-p)|U_{u,s}|) \left| B_{u,2s-1}, U_{u,s} \right. \right) \leq 2 \exp \left( -\frac{1}{3}\delta^2 \left( [n] \setminus B_{u,2s-1} \right)(1 - (1-p)|U_{u,s}|) \right). \]  

(7.6)

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

\[ \Pr \left( |U_{u,s}| \notin (1 \pm \delta)\left( [n/2] \setminus B_{u,2(s-1)} \right)(1 - (1-p)|S_{u,s-1}|) \left| B_{u,2(s-1)}, S_{u,s-1} \right. \right) \leq 2 \exp \left( -\frac{1}{3}\delta^2 \left( [n/2] \setminus B_{u,2(s-1)} \right)(1 - (1-p)|S_{u,s-1}|) \right). \]  

(7.7)

Let us define the events

\[ A_{u,s}^1(\delta) = \left\{ |S_{u,s}| \in (1 \pm \delta)([n] \setminus B_{u,2s-1})(1 - (1-p)|U_{u,s}|) \right\}, \]  

(7.8)

\[ A_{u,s}^2(\delta) = \left\{ |U_{u,s}| \in (1 \pm \delta)\left( [n/2] \setminus B_{u,2(s-1)} \right)(1 - (1-p)|S_{u,s-1}|) \right\}. \]  

(7.9)

Since \( p \in (0,1) \) and hence \( 1 - (1-p)x \leq px \) for all \( x \geq 1 \), we have that under events \( A_{u,s}^1(\delta) \cap A_{u,s}^2(\delta) \),

\[ |S_{u,s}| \leq (1 + \delta)np|U_{u,s}| \quad \text{and} \quad |U_{u,s}| \leq (1 + \delta)\left( \frac{n/2}{2} \right)p|S_{u,s-1}|, \]

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

\[ |S_{u,s}| \leq \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^s = (1 + \delta)^{2s^2 - 3s^3} n^{2\kappa s}, \]  

(7.10)
and

\[ |\mathcal{U}_{u,s}| \leq (1 + \delta) \frac{pm^2}{8} \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^{s-1} = (1 + \delta)^{2s-1} 2^{-3s} n^{\frac{1}{2} + \kappa (2s-1)} . \tag{7.11} \]

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

\[ |\mathcal{B}_{u,2s}| \leq 1 + \sum_{\ell=1}^{s} (2|\mathcal{U}_{u,\ell}| + |\mathcal{S}_{u,\ell}|) \]

\[ \leq 1 + \sum_{\ell=1}^{s} \left( 2(1 + \delta) \frac{pm^2}{8} \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^{\ell-1} + \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^{\ell} \right) \]

\[ = 1 + \left( 2(1 + \delta) \frac{pm^2}{8} + \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right) \right) \sum_{\ell=1}^{s} \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^{\ell-1} \]

\[ = O(pm^2(p^2 n^3)^{s-1}) = O(n^{\kappa (2s-1) + \frac{1}{2}}) = O(n^{1-\kappa}) = o(n). \tag{7.12} \]

With a similar argument we can show that

\[ \sum_{\ell=0}^{t} |\mathcal{S}_{u,\ell}| \leq \sum_{\ell=0}^{t} \left( (1 + \delta)^2 \frac{p^2 n^3}{8} \right)^{\ell} = O((p^2 n^3)^t) = O(n^{2\kappa t}) = o(n). \tag{7.13} \]

The last step follows from checking that when \( \kappa \in \left[ \frac{1}{4}, \frac{1}{2} \right) \), \( t = 1 \) such that \( n^{2\kappa t} = o(n) \), and when \( \kappa \in (0, \frac{1}{4}) \), from \( t \leq \frac{1}{4\kappa} + 1 \), 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 \( \bigcup_{\ell=1}^{t} \mathcal{U}_{u,\ell} \subset \mathcal{V}_A \), 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 (7.13),

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

Using (7.12), we establish lower bounds on \( |\mathcal{S}_{u,s}| \) and \( |\mathcal{U}_{u,s}| \) next. Note that, for \( p \in (0, 1) \), \( 1 - p \leq e^{-p} \) and for any \( x \in (0, 1) \), \( e^{-x} \leq 1 - x + x^2 \). It follows that \( 1 - (1 - p)^x \geq px(1 - px) \). For \( s \in [t] \) we can show that

\[ |\mathcal{U}_{u,s}| \geq (1 - \delta) \frac{(n(1 - o(1)))^2(1 - o(1))}{8} (1 - (1 - p)^{|\mathcal{S}_{u,s-1}|}) \]

\[ \geq (1 - \delta) \frac{n^2}{8} p|\mathcal{S}_{u,s-1}|(1 - p)^{|\mathcal{S}_{u,s-1}|}(1 - o(1)) \]

\[ = (1 - \delta) \frac{n^2}{8} p|\mathcal{S}_{u,s-1}|(1 - o(1)). \]

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

\[ |\mathcal{S}_{u,s}| \geq (1 - \delta)n(1 - o(1))(1 - (1 - p)^{|\mathcal{U}_{u,s}|}) \]

\[ \geq (1 - \delta)n(1 - o(1))p|\mathcal{U}_{u,s}|(1 - p|\mathcal{U}_{u,s}|) \]

\[ \geq (1 - \delta)n(1 - o(1))p|\mathcal{U}_{u,s}|(1 - o(1)) \]

\[ = (1 - \delta)pn|\mathcal{U}_{u,s}|(1 - o(1)), \]
and for \( s = t \),

\[
|S_{u,t}| \geq (1 - \delta) \frac{n}{2} (1 - o(1)) (1 - (1 - p)|U_{u,t}|).
\]

Then for \( s \in [t] \),

\[
|U_{u,s}| \geq (1 - \delta) \frac{p^2 n^3}{8} |U_{u,s-1}| (1 - o(1))
\]

\[
\geq \left( 1 - \delta \right)^{2} \frac{p^2 n^3}{8} \frac{s - 1}{s} |U_{u,1}| (1 - o(1))
\]

\[
\geq \left( 1 - \delta \right)^{2} \frac{p^2 n^3}{8} \left( 1 - \delta \right)^{\frac{p n^2}{8}} (1 - o(1))
\]

\[
= \left( 1 - \delta \right)^{2s - 1} 2^{-3s} n^\frac{1}{2} + \kappa (2s - 1) (1 - o(1));
\]

\( (7.14) \)

for \( s \in [t - 1] \),

\[
|S_{u,s}| \geq (1 - \delta) p n^2 \frac{n}{8} |S_{u,s-1}| (1 - o(1)) \geq \left( 1 - \delta \right)^{2} \frac{p^2 n^3}{8} (1 - o(1)),
\]

\[
= \left( 1 - \delta \right)^{2s - 2} 2^{-3s} n^{2\kappa} (1 - o(1));
\]

\( (7.15) \)

and for \( s = t, |S_{u,t}| \geq \left( 1 - \delta \right)^{2t - 3t - 1} n^{2\kappa} (1 - o(1)). \)

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

\[
\mathbb{P} \left( \overline{\cap_{s=1}^{t} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))} \right) = \mathbb{P} \left( \cup_{s=1}^{t} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta)) \right)
\]

\[
= \sum_{s=1}^{t} \mathbb{P} \left( \overline{(A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))} \right) \cap \cap_{s=1}^{s-1} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))
\]

\[
\leq \sum_{s=1}^{t} \mathbb{P} \left( \overline{(A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))} \right) \cap \cap_{s=1}^{s-1} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))
\]

\[
\leq \sum_{s=1}^{t} \mathbb{P} \left( \overline{(A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))} \right) \cap \cap_{s=1}^{s-1} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))
\]

\[
\leq \sum_{s=1}^{t} \mathbb{P} \left( A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta) \right) \cap \cap_{s=1}^{s-1} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))
\]

We bound the each of the two summation terms on the right hand side in the last inequality next. Using \( (7.6) \) and \( (7.10) \), we have

\[
\sum_{s=1}^{t} \mathbb{P} \left( \overline{(A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))} \right) \cap \cap_{s=1}^{s-1} (A_{u,s}^1 (\delta) \cap A_{u,s}^2 (\delta))
\]

\[
\leq \sum_{s=1}^{t-1} 2 \exp \left( -\frac{1}{3} \delta^2 (1 - \delta) \frac{p^2 n^3}{2} \left( 1 - \delta \right)^{2} \frac{p^2 n^3}{2} \right) \frac{s - 1}{s} (1 - o(1)) = O \left( \exp \left( -\Theta(n^{2\kappa}) \right) \right).
\]
Similarly, using (7.7) and (7.11), we have

\[
\sum_{s=1}^{t} P\left( -A_{u,s}^2(\delta) \mid A_{u,s}^1(\delta) \cap A_{u,h}^2(\delta) \right) \\
\leq \sum_{s=1}^{t+1} 2 \exp\left( -\frac{1}{3} \delta^2 \frac{n^2 p}{2} \left( (1 - \delta)^2 \frac{n^2 p^2 n^3}{2} \right)^{s-1} (1 - o(1)) \right) \\
\leq 4 \exp\left( -\frac{1}{3} \delta^2 \frac{n^2 p}{2} (1 - o(1)) \right) = O\left( \exp\left( -\Theta(n^\frac{1}{2} + \kappa) \right) \right).
\]

Putting it all together, we have that

\[
P\left( -\cap_{s=1}^{t}(A_{u,s}^1(\delta) \cap A_{u,s}^2(\delta)) \right) \leq O\left( \exp\left( -\Theta(n^{2\kappa}) \right) \right) + O\left( \exp\left( -\Theta(n^\frac{1}{2} + \kappa) \right) \right)
= O\left( \exp\left( -\Theta(n^{2\kappa}) \right) \right),
\]

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.

\[\square\]

### 7.2 Concentration of Quadratic Form One

Let \( A_{u,t}^3(\delta) \) denote the event that (7.1) holds for all \( s \in [t-1] \), (7.2) holds, (7.3) holds for all \( s \in [t] \), and (7.4) holds. Lemma 7.1 established that this event holds with high probability. Conditioned on the event \( A_{u,t}^3(\delta) \), 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 (4.1), \( \delta \in (0, \frac{1}{2}) \), and \( \psi \in (0, \kappa) \). For any \( u \in [n] \), with probability \( 1 - 2 \exp( -n^{2\psi}(1 - o(1))) \),

\[ |e_u^T Q \hat{N}_{u,t} - e_u^T A_{u,t} = \frac{16 \lambda_k^2 t - 2 n \psi}{(1 - \delta)n^\kappa}. \]

**Proof.** Recall that conditioning on event \( A_{u,t}^3(\delta) \) simply imposes the restriction that the neighborhood of \( u \in [n] \) grows at a specific rate. This event is independent from latent parameters \( \{a \}_{a \in [n]} \), the precise entries in \( \Omega_1 \) as well as associated values, i.e. \( M_1 \).

Conditioned on \( A_{u,t}^3(\delta) \), let \( F_{u,s} \) 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 \( S_{u,h'} \) for \( h' \leq \lceil s/2 \rceil \), \( U_{u,h''} \) for \( h'' \leq \lceil s/2 \rceil \), associated latent parameters as well as edges of \( \Omega_1 \). Specifically, \( F_{u,0} \) contains information about latent parameter \( \theta_a \) associated with \( a \in [n] \); \( F_{u,s} \) contains information about latent parameters \( \cup_{h=1}^{\lceil s/2 \rceil} \{ \theta_{a} \}_{a \in S_{u,h}} \cup_{h=1}^{\lceil s/2 \rceil} \{ \theta_b, \theta_c \}_{(b,c) \in U_{u,h}} \), and all the associated edges and observations. This implies that \( F_{u,0} \subset F_{u,1} \subset F_{u,2} \), 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 \( Q \) denote the \( r \times n/2 \) matrix where \( Q_{kb} = q_k(\theta_{b_1})q_k(\theta_{b_2}) \) 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 \( e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^{2t} Q e_u \). For \( 1 \leq s \leq 2t \), define

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

\[
D_{u,s} = Y_{u,s} - Y_{u,s-1},
\]

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

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 \( e_k^T \Lambda^{2t-s} Q \tilde{W}_{u,(s+1)/2} \) 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 \( 1 \leq s \leq t \), i.e.

\[
\mathbb{E}[D_{u,s} \mid \mathcal{F}_{u,s-1}] = 0.
\] (7.16)

Furthermore, for properly chosen \( \nu_s \) as specified in Lemmas 7.4 and 7.5,

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

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

We can then apply Proposition 7.3 with any arbitrarily small \( \alpha_s \) such that for any \( x \geq 0 \),

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

where for \( n \) large enough,

\[
\sum_{s=1}^{2t} \nu_s^2 \leq \sum_{s=1}^{t} \frac{(1 + 4\pi)\lambda_k^{4t-4s}2^{3s-1}(1 + o(1))}{(1 - \delta)^{2s}n^{2s}} + \sum_{s=1}^{t} \frac{(1 + 16\pi)72\lambda_k^{4t-4s+2}B^42^{3s-1}(1 + o(1))}{(1 - \delta)^{2s-1}n^{\min\{1, \frac{1}{2} + \kappa(2s-1)\}}}
\]

\[
\leq \frac{8(1 + 4\pi)\lambda_k^{4t-4}(1 + o(1))}{(1 - \delta)^{2t}n^{2\kappa}},
\]

and \( 1 + 4\pi \leq 16 \).

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 \( 1 - 2 \exp(-n^2\psi(1 - o(1))) \),

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

\( \square \)

We recall the following concentration inequality for Martingale difference sequence, cf. [32, Theorem 2.19]:

**Proposition 7.3.** Let \( \{D_k, \mathcal{F}_k\}_{k \geq 1} \) be a martingale difference sequence such that \( \mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2 / 2} \) almost surely for all \( \lambda \in \mathbb{R} \). Then for all \( x > 0 \),

\[
\mathbb{P}\left( \sum_{k=1}^{n} D_k \geq x \right) \leq 2 \exp\left(-\frac{x^2}{2 \sum_{k=1}^{n} \nu_k^2}\right).
\] (7.17)
Lemma 7.4. For any $s \in [t]$,\[ \mathbb{E} \left[ D_{u,2s} \mid \mathcal{F}_{u,2s-1}, \mathcal{A}_{u,t}^3(\delta) \right] = 0. \]

Let $\nu = \sqrt{\frac{Q(1+4\pi)}{2}}$, and\[ Q = \frac{B^2 \lambda_k 2^{t-4} s^3 (1 + o(1))}{(1 - \delta)^{2s^2 n^2 k^2 s}}. \]

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

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 $B_{u,2s-1} = \bigcup_{l=1}^{s} \bigcup_{h=1}^{e} S_{u,l}$, the vertex latent variables $\{\theta_i\}_{i \in S_{u,2s-1}}$ and the edges and corresponding weights. Let us additionally condition on the set $S_{u,s}$. As $2s$ is even,
\[ D_{u,2s} = Y_{u,2s} - Y_{u,2s-1} \]
\[ = \lambda_k^{2t-2} \left( e_k^T Q \tilde{N}_{u,s} - \lambda_k e_k^T Q \tilde{W}_{u,s} \right) \]
\[ = \lambda_k^{2t-2} \left( \frac{1}{|S_{u,s}|} \sum_{i \in [n]} N_{u,s}(i) q_k(\theta_i) - \lambda_k e_k^T Q \tilde{W}_{u,s} \right) \]
\[ = \lambda_k^{2t-2} \left( \frac{1}{|S_{u,s}|} \sum_{i \in S_{u,s}} \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} W_{u,s}(a) 1_{(a=\pi(i))} M_1(a_1,a_2,i) q_k(\theta_i) - \lambda_k e_k^T Q \tilde{W}_{u,s} \right). \]

Let us define
\[ X_i = \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} W_{u,s}(a) 1_{(a=\pi(i))} M_1(a_1,a_2,i) q_k(\theta_i) \]
\[ = \sum_{a=(a_1,a_2) \in \mathcal{U}_{u,s}} W_{u,s}(a) 1_{(a=\pi(i))} (f(\theta_{a_1}, \theta_{a_2}, \theta_i) + \epsilon_{a_1,a_2}) q_k(\theta_i). \]

The randomness in $X_i$ only depends on $\theta_i, \epsilon_{a_1,a_2}, 1_{(a=\pi(i))}$. Note that we already conditioned on $\theta_{a_1}, \theta_{a_2}$ for $a \in \mathcal{U}_{u,s} \subset B_{2s-1}$. $X_i$ 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 $i \in S_{u,s}$). For any vertex $i \in S_{u,s}$, 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 \( a \in U_{u,s} \),

\[
\mathbb{E}[X_i \mid i \in S_{u,s}] = \mathbb{E} \left[ \sum_{a=(a_1,a_2) \in U_{u,s}} W_{u,s}(a) \mathbb{I}_{(a=\pi(i))}(f(\theta_{a_1}, \theta_{a_2}, \theta_i) + \epsilon_{a_1a_2i}) q_k(\theta_i) \right] = \mathbb{E} \left[ \sum_{a=(a_1,a_2) \in U_{u,s}} \frac{1}{|U_{u,s}|} \mathbb{E} \left[ W_{u,s}(a) \sum_h \lambda_h q_h(\theta_{a_1}) q_h(\theta_{a_2}) q_h(\theta_i) q_k(\theta_i) \right] \right] = \sum_{a=(a_1,a_2) \in U_{u,s}} \hat{W}_{u,s}(a) \lambda_k q_k(\theta_{a_1}) q_k(\theta_{a_2}) = e_k^T \Lambda \tilde{Q} \hat{W}_{u,s}.
\]

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

\[
\mathbb{P} \left( |D_{u,2s}| \geq z \mid \mathcal{F}_{u,2s-1}, S_{u,s} \right) \leq 2 \exp \left( -\frac{2 |S_{u,s}| z^2}{\lambda_k^{4-4s} B^2} \right).
\]

If we condition on the event \( \mathcal{A}_{u,t}^3(\delta), |S_{u,s}| \geq (1-\delta)^{2s-3s-1} n^{2\kappa s} (1-o(1)) \) for \( s \in [t] \). Therefore,

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

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

\[\square\]

**Lemma 7.5.** For any \( s \in [t] \),

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

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

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

For any \( \lambda \in \mathbb{R} \),

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

**Proof.** As \( 2s-1 \) is odd,

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

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

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

30
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

\[
e_k^T \mathcal{Q} \hat{W}_{u,s} = \frac{1}{|\mathcal{U}_{u,s}|} \sum_{i = (i_1,i_2) \in \mathcal{U}_{u,s}} W_{u,s}(i) q_k(\theta_{i_1}) q_k(\theta_{i_2})
\]

\[
= \frac{1}{|\mathcal{U}_{u,s}|} \sum_{i = (i_1,i_2) \in \mathcal{U}_{u,s}} \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \mathbb{I}_{\{v = \pi(i)\}} M_1(v,i) q_k(\theta_{i_1}) q_k(\theta_{i_2})
\]

\[
= \frac{1}{|\mathcal{U}_{u,s}|} \sum_{i = (i_1,i_2) \in \mathcal{U}_{u,s}} X_i,
\]

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

\[
X_i = \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \mathbb{I}_{\{v = \pi(i)\}} M_1(v,i) q_k(\theta_{i_1}) q_k(\theta_{i_2})
\]

\[
= \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \mathbb{I}_{\{v = \pi(i)\}} \left( \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) + \epsilon_{v,i_1,i_2} \right) q_k(\theta_{i_1}) q_k(\theta_{i_2}).
\]

Conditioned on \( \mathcal{F}_{u,2(s-1)} \) the randomness in \( X_i \) only depends on \( \theta_{i_1}, \theta_{i_2}, \epsilon_{\pi(i)i_1,i_2} \), 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 \( \epsilon_{\pi(i)i_1,i_2} \) 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 \( v \in \mathcal{S}_{u,s-1} \),

\[
\mathbb{E}[X_i \mid \theta_{i_1}, \theta_{i_2}, i \in \mathcal{U}_{u,s}]
\]

\[
= \mathbb{E} \left[ \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \mathbb{I}_{\{v = \pi(i)\}} \left( \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) \right) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \mid \theta_{i_1}, \theta_{i_2} \right]
\]

\[
= \frac{1}{|\mathcal{S}_{u,s-1}|} \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \mathbb{I}_{\{v = \pi(i)\}} \left( \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) \right) q_k(\theta_{i_1}) q_k(\theta_{i_2}).
\]

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

\[
P \left( \|e_k^T \mathcal{Q} \hat{W}_{u,s} - \mathbb{E}[e_k^T \mathcal{Q} \hat{W}_{u,s} \mid \{ \theta_{i_1}, \theta_{i_2} \}_{i \in \mathcal{U}_{u,s}, \mathcal{U}_{u,s}}] \| > z \mid \mathcal{F}_{u,2(s-1)}, \mathcal{U}_{u,s}, \{ \theta_{i_1}, \theta_{i_2} \}_{i \in \mathcal{U}_{u,s}} \right)
\]

\[
\leq 2 \exp \left( -\frac{|\mathcal{U}_{u,s}| z^2}{B^4} \right). \tag{7.18}
\]

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,

\[
\frac{1}{|\mathcal{U}_{u,s}|} \sum_{i \in \mathcal{U}_{u,s}} \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2})
\]

\[
\approx \frac{1}{|\mathcal{V}_A \setminus \mathcal{B}_{u,2(s-1)}|} \sum_{i \in \mathcal{V}_A \setminus \mathcal{B}_{u,2(s-1)}} \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}).
\]
Next, we formalize it. To that end, conditioned on the size $|U_{u,s}|$, the set $U_{u,s}$ is a uniform random sample of the possible set of vertices $V_A \setminus B_{u,2(s-1)}$. The above expression on the left is thus the mean of a random sample $U_{u,s}$ without replacement from $V_A \setminus B_{u,2(s-1)}$. 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. [16, Theorem 4]. Therefore, using
\[
\sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) = |f(\theta_v, \theta_{i_1}, \theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2})| \leq B^2,
\]
we can apply Hoeffding’s inequality to argue that
\[
P\left(\frac{1}{|U_{u,s}|} \sum_{i \in U_{u,s}} q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) - \frac{1}{|V_A \setminus B_{u,2(s-1)}|} \sum_{i \in V_A \setminus B_{u,2(s-1)}} q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \geq z \mid \{\theta_{i_1}, \theta_{i_2}\} \in V_A, |U_{u,s}|\right) \leq 2 \exp \left(-\frac{|U_{u,s}| z^2}{B^4}\right).
\] (7.19)

Finally, we need to account for the randomness in $\{\theta_{i_1}, \theta_{i_2}\} \in V_A$, arguing that with high probability
\[
\frac{1}{|V_A \setminus B_{u,2(s-1)}|} \sum_{i \in V_A \setminus B_{u,2(s-1)}} \sum_l \lambda_l q_l(\theta_v) q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) \approx \lambda_k q_k(\theta_v).
\]
To formalize this, we start by recalling that
\[
V_A \setminus B_{u,2(s-1)} = \{(i_1, i_2) \text{ s.t. } i_1 < i_2 \text{ and } \{i_1, i_2\} \subset [n/2] \setminus B_{u,2(s-1)}\}.
\]
Let $n_{u,s} = |[n/2] \setminus B_{u,2(s-1)}|$, then $|V_A \setminus B_{u,2(s-1)}| = \binom{n_{u,s}}{2}$. Then the above summation can be written as a pairwise U-statistic,
\[
U = \frac{1}{|V_A \setminus B_{u,2(s-1)}|} \sum_{(i_1, i_2)} g(\theta_{i_1}, \theta_{i_2})
\]
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, $E[\sum_l \lambda_l q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2})] = \lambda_k q_k(\theta_v)$ by the orthogonality model assumption. Therefore, by Lemma 4
\[
P\left(\frac{1}{|V_A \setminus B_{u,2(s-1)}|} \sum_{i \in V_A \setminus B_{u,2(s-1)}} \sum_l \lambda_l q_l(\theta_{i_1}) q_l(\theta_{i_2}) q_k(\theta_{i_1}) q_k(\theta_{i_2}) - \lambda_k q_k(\theta_v) \geq z \right) \leq 2 \exp \left(-\frac{n_{u,s} z^2}{8B^4}\right).
\] (7.20)

By putting together all calculations, it also follows that
\[
E[e_k^T Q \tilde{W}_{u,s}] = e_k^T A \tilde{Q} \tilde{N}_{u,s-1},
\]
and for $z_1, z_2, z_3 > 0$, with probability at least
\[
1 - 2 \exp \left(-\frac{|U_{u,s}| z_1^2}{B^4}\right) - 2 \exp \left(-\frac{|U_{u,s}| z_2^2}{B^4}\right) - 2 \exp \left(-\frac{n_{u,s} z_3^2}{8B^4}\right).
\]
it holds that
\[
|e_k^T Q \tilde{W}_{u,s} - e_k^T \Lambda Q \tilde{N}_{u,s-1}| \\
\leq |e_k^T Q \tilde{W}_{u,s} - \mathbb{E}[e_k^T Q \tilde{W}_{u,s} \mid \{\theta_{i_1}, \theta_{i_2}\} \in \mathcal{U}_{u,s}, \mathcal{U}_{u,s}]| \\
+ \frac{1}{|\mathcal{S}_{u,s-1}|} \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \left( \frac{1}{|\mathcal{U}_{u,s}|} \sum_{i \in \mathcal{U}_{u,s}} \lambda_i q_i(\theta_v) q_i(\theta_{i_1}) q_i(\theta_{i_2}) q_i(\theta_{i_3}) q_i(\theta_{i_4}) \right) \\
- \frac{1}{|\mathcal{V}_A \setminus B_{u,2(s-1)}|} \sum_{i \in \mathcal{V}_A \setminus B_{u,2(s-1)}} \lambda_i q_i(\theta_v) q_i(\theta_{i_1}) q_i(\theta_{i_2}) q_i(\theta_{i_3}) q_i(\theta_{i_4}) \\
+ \frac{1}{|\mathcal{S}_{u,s-1}|} \sum_{v \in \mathcal{S}_{u,s-1}} N_{u,s-1}(v) \left( \frac{\sum_{i \in \mathcal{V}_A \setminus B_{u,2(s-1)}} \lambda_i q_i(\theta_v) q_i(\theta_{i_1}) q_i(\theta_{i_2}) q_i(\theta_{i_3}) q_i(\theta_{i_4})}{|\mathcal{V}_A \setminus B_{u,2(s-1)}|} \right) \\
\leq z_1 + \frac{1}{|\mathcal{S}_{u,s-1}|} \sum_{v \in \mathcal{S}_{u,s-1}} |N_{u,s-1}(v)| z_2 + \frac{1}{|\mathcal{S}_{u,s-1}|} \sum_{v \in \mathcal{S}_{u,s-1}} |N_{u,s-1}(v)| z_3
\]

since \(\|N_{u,s-1}\|_\infty \leq 1\). Conditioned on \(\mathcal{A}_{u,t}^3(\delta), n_{u,s} = n/2(1 - o(1))\), and
\[
|\mathcal{U}_{u,s}| \in \left[ (1 - \delta)^{2s-1/2} - 3n^{1/2 + \kappa(2s-1)}(1 - o(1)), (1 + \delta)^{2s-1/2} - 3n^{1/2 + \kappa(2s-1)} \right].
\]

As a result, for \(z_1 = z_2 = z_3\), the expression in (7.18) and (7.19) asymptotically dominate the expression in (7.20). It follows that, with appropriate choice of \(z_1 = z_2 = z_3\) in the above,
\[
\mathbb{P}(|D_{u,2s-1}| \geq z \mid \mathcal{F}_{2s-2}, \mathcal{A}_{u,t}^3(\delta)) \\
\leq 6 \exp \left( - \frac{(1 - \delta)^{2s-1/2} - 3n^{\min\{1, 1/2 + \kappa(2s-1)\}}(1 - o(1))z^2}{72\lambda_k^{4s-4s+2} B^4} \right)
\]

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

\[\square\]

### 7.3 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 (4.1), i.e. \(t = \left[ \frac{1}{\delta_{\max}^2 n^\psi} \right], \delta \in (0, 1)\), and
\[
x = \frac{16\lambda_k^{4t - 2} n^\psi}{(1 - \delta)n^\gamma}
\]
define
\[
\mathcal{A}_{u,k,t}^4(x, \delta) = \left\{ |e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^2 Q e_u| < x \right\} \cap \mathcal{A}_{u,t}^3(\delta).
\]

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 \(\kappa \in (0, 1/2)\), \(t\) as defined in (4.1), and \(\delta \in (0, 1/2)\). For any \(u, v \in [n]\), conditioned on \(\cap_{k=1}^r (\mathcal{A}_{u,k,t}^4(x, \delta) \cap \mathcal{A}_{v,k,t}^4(x, \delta))\),
\[
|\tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t} - e_u^T Q^T \Lambda^{2(t+1)} Q e_v| \leq x^2 \left( \sum_{k=1}^r \lambda_k^2 \right) + xB \left( \sum_{k=1}^r 2\lambda_k^{2(t+1)} \right).
\]

33
Proof. Proof of Lemma 7.6. Assuming event \( \cap_{k=1}^r (A_{u,k,t}^4(x, \delta) \cap A_{v,k,t}^4(x, \delta)) \) holds,

\[
\begin{align*}
\left| \tilde{N}_{u,t}^T \Lambda^2 Q \tilde{N}_{v,t} - e_u^T Q^T \Lambda^2(2t+1) Q e_v \right| \\
\leq \left| (\tilde{N}_{u,t}^T \Lambda^2 - e_u^T Q^T \Lambda^2) (\Lambda^2 Q \tilde{N}_{v,t} - \Lambda^2(2t+1) Q e_v) \right| \\
+ \left| (\tilde{N}_{u,t}^T \Lambda^2 - e_u^T Q^T \Lambda^2) \Lambda^2(2t+1) Q e_v \right| + \left| e_u^T Q^T \Lambda^2(2t+1) (Q \tilde{N}_{v,t} - \Lambda^2 Q e_v) \right| \\
\leq \sum_{k=1}^r (e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^2 Q e_u) (e_k^T Q \tilde{N}_{v,t} - e_k^T \Lambda^2(2t+1) Q e_v) \\
+ \sum_{k=1}^r (e_k^T Q \tilde{N}_{u,t} - e_k^T \Lambda^2 Q e_u) e_k^T \Lambda^2 Q e_v \\
+ \sum_{k=1}^r (e_k^T \Lambda^2(2t+1) Q e_v) (e_k^T Q \tilde{N}_{v,t} - e_k^T \Lambda^2 Q e_v). \quad (7.21)
\end{align*}
\]

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

\[
\left| \tilde{N}_{u,t}^T \Lambda^2 Q \tilde{N}_{v,t} - e_u^T Q^T \Lambda^2(2t+1) Q e_v \right| \leq x^2 \left( \sum_{k=1}^r \lambda_k^2 \right) + xB \left( \sum_{k=1}^r 2\lambda_k^2(2t+1) \right). \quad (7.22)
\]

\[\Box\]

### 7.4 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

\[
A'_{u,v,t}(x, \delta) = \cap_{k=1}^r (A_{u,k,t}^4(x, \delta) \cap A_{v,k,t}^4(x, \delta)). \quad (7.23)
\]

**Lemma 7.7.** Let \( p = n^{-3/2+\kappa} \) for \( \kappa \in (0, \frac{1}{2}) \), \( t \) as defined in (4.1), \( \delta \in (0, \frac{1}{2}) \), and

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

Let \( S \equiv S_{u,v,t} = [n] \setminus (B_{u,2t} \cup B_{v,2t} \cup \lfloor n/2 \rfloor) \). Then, under event \( A'_{u,v,t}(x, \delta) \),

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

with probability at least \( 1 - 4 \exp(-n^2(1-o(1))) - O(n^{-6}) \) with \( \psi \in (0, \kappa) \).
Proof. First, note that $A_{u,v,t}^r(x,\delta)$ includes events $A_{u,t}^3(\delta)$ and $A_{v,t}^3(\delta)$. This implies that $|S| = \frac{t}{2} - o(n) = \frac{n(1-o(1))}{2}$. Furthermore, it implies that $|S_{u,t}|$ and $|S_{v,t}|$ are both greater than or equal to $(1 - o(2t-3t^{-1}n^{2\kappa t}(1-o(1)))$. As a result,

$$
|S|^2|\mathcal{P}|S_{u,t}|S_{v,t}| \geq \frac{n^2}{8}p^2\left(\frac{(1 - \delta)^2}{8}n^{2\kappa}\right)^{2t}(1 - o(1))
$$

$$
\geq n^{2+4\kappa t}n^{2(\frac{\delta}{2} + \kappa)}\frac{1}{8}\left(\frac{(1 - \delta)^2}{8}\right)^{2t}(1 - o(1))
$$

$$
= \Theta(n^{-1+2\kappa(2t+1)})
$$

$$
= \Omega(n^{2\kappa}).
$$

(7.24)

(7.25)

(7.26)

(7.27)

The asymptotic relationships follow from the choice of $t \geq \frac{1}{16\kappa}$, and the fact that $\delta$ 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 $\Omega_1$, the observation noise terms $\epsilon_{a\alpha\beta}$, and the latent variables $\theta_a, \theta_\alpha, \theta_\beta$. Since we enforce that $\alpha$ and $\beta$ 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

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

for $\phi = [16/(1 - 2\kappa)]$, 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 $a \notin S_{u,t} \subseteq B_{u,2t}$ and $N_{v,t}(b) = 0$ if $b \notin S_{v,t} \subseteq B_{v,2t}$. Further, conditioned on event $A_{u,v,t}^r(x, \delta)$, all the information associated with $\mathcal{B}_{u,2t}$ and $\mathcal{B}_{v,2t}$ is revealed; however, information about $[n] \setminus (\mathcal{B}_{u,2t} \cup \mathcal{B}_{v,2t})$ is not. Let $\mathcal{F}(u, v, t, x, \delta)$ denote all the information revealed such that event $A_{u,v,t}^r(x, \delta)$ 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 $\Omega_1$) and the observation noise are independent for distinct pairs $(\alpha, \beta)$ and $(\alpha', \beta')$. 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 $T(\alpha, \beta)$, it follows that (using notation $\mathcal{F} = \mathcal{F}(u, v, t, x, \delta)$)

$$
\text{Var} \left[ \sum_{\alpha < \beta \in S \times S} T(\alpha, \beta) \mid \mathcal{F}, \{\theta_i\}_{i \in [n]} \right] = \sum_{\alpha < \beta \in S \times S} \text{Var} \left[ T(\alpha, \beta) \mid \mathcal{F}, \{\theta_i\}_{i \in [n]} \right] \leq 2 \left( \frac{|S|}{2} \right)^2|S_{u,t}|S_{v,t}|(1 + o(1)).
$$

35
The variables $\tilde{T}(\alpha, \beta)$ are also independent across $(\alpha, \beta)$ 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 $z = 2n^\psi \left( \binom{|S|}{2}p^2 |S_{u,t}| |S_{v,t}| \right)^{1/2}$ for $\psi \in (0, \kappa)$, it follows that

$$\mathbb{P}\left( \left| \sum_{\alpha < \beta \in S \times S} (\tilde{T}(\alpha, \beta) - \mathbb{E}[\tilde{T}(\alpha, \beta) \mid F, \{\theta_i\}_{i \in [n]}]) \right| \geq z \mid F, \{\theta_i\}_{i \in [n]} \right) \leq 2 \exp \left( - \frac{z^2}{2 \left( \binom{|S|}{2}p^2 |S_{u,t}| |S_{v,t}| (1 + o(1)) + \frac{2\phi^2}{3} \right)} \right) = 2 \exp(-n^2\phi(1-o(1))).$$

(7.29)

The last equality arises from the observation that $t$ is chosen such that conditioned on $F$, we can plug in (7.27) to show that for our choice of $z$, it holds that

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

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

$$|\mathbb{E}[T(\alpha, \beta) \mid F] - \mathbb{E}[\tilde{T}(\alpha, \beta) \mid F]| \leq \mathbb{E}[|T(\alpha, \beta) - \tilde{T}(\alpha, \beta)| \mid F] = \mathbb{E}[|T(\alpha, \beta)| - \min(|T(\alpha, \beta)|, \phi^2) \mid F] = \mathbb{E}[|I_{|T(\alpha, \beta)| \geq \phi^2}(|T(\alpha, \beta)| - \phi^2) \mid F].$$

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

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

(7.30) (7.31) (7.32) (7.33)

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

By plugging in our choice of $\phi$ into Lemma 7.9, it also follows that

$$\mathbb{P}\left( \cup_{\alpha, \beta} \{\tilde{T}(\alpha, \beta) \neq T(\alpha, \beta)\} \mid F \right) \leq O(n^{-6}).$$

By combining (7.29), (7.33), and (7.34), with probability at least $1 - 2 \exp(-n^2\psi(1-o(1))) = O(n^{-6})$,

$$\sum_{\alpha < \beta \in S \times S} \left| T(\alpha, \beta) - \mathbb{E}[T(\alpha, \beta) \mid F, \{\theta_i\}_{i \in [n]}] \right| \leq 2n^\psi \left( \binom{|S|}{2}p^2 |S_{u,t}| |S_{v,t}| \right)^{1/2} + O(n^{-6}),$$

(7.34) (7.35)
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,

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

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

\[
g(\theta_\alpha, \theta_\beta) = \frac{1}{p^2 |S_u,t| |S_v,t|} \mathbb{E} \left[ T(\alpha, \beta) \middle| \mathcal{F}, \{\theta_i\}_{i \in [n]} \right]
\]

\[
= \frac{1}{p^2 |S_u,t| |S_v,t|} \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) \times \mathbb{E} \left[ M_B(a, (\alpha, \beta)) M_B(b, (\alpha, \beta)) \middle| \mathcal{F}, \{\theta_i\}_{i \in [n]} \right]
\]

\[
= \frac{1}{|S_u,t| |S_v,t|} \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) F(a, \alpha, \beta) F(b, \alpha, \beta).
\]

It follows by boundedness of entries in \( F \) and the fact that \( \|N_{u,t}\|_\infty \leq 1 \) and \( \|N_{u,t}\|_0 = |S_{u,t}| \), that \( |g(\theta_\alpha, \theta_\beta)| \leq 1 \) almost surely. Therefore, by Lemma 4 and choosing \( z = \sqrt{8n^2} |S|^{-1/2} \),

\[
\mathbb{P} \left( \left| \frac{1}{\binom{|S|}{2} p^2 |S_u,t| |S_v,t|} \sum_{\alpha < \beta \in S \times S} \left( \mathbb{E} \left[ T(\alpha, \beta) \middle| \mathcal{F}, \{\theta_i\}_{i \in [n]} \right] - \mathbb{E} \left[ T(\alpha, \beta) \middle| \mathcal{F} \right] \right) \right| \geq z \right) \geq 2 \exp \left( - \frac{|S| z^2}{8} \right) = 2 \exp(n^2 \psi).
\]

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

\[
\frac{1}{\binom{|S|}{2} p^2 |S_u,t| |S_v,t|} \mathbb{E} \left[ \sum_{\alpha < \beta \in S \times S} T(\alpha, \beta) \middle| \mathcal{F} \right]
\]

\[
= \frac{1}{\binom{|S|}{2}} \sum_{\alpha < \beta \in S \times S} \frac{1}{|S_u,t| |S_v,t|} \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) \mathbb{E}[F(a, \alpha, \beta) F(b, \alpha, \beta)]
\]

\[
= \frac{1}{\binom{|S|}{2}} \sum_{\alpha < \beta \in S \times S} \frac{1}{|S_u,t| |S_v,t|} \sum_{a \neq b \in [n]} N_{u,t}(a) N_{v,t}(b) \sum_k \lambda_k^2 q_k(\theta_a) q_k(\theta_b)
\]

\[
= \tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t} - \sum_{a \in [n]} \tilde{N}_{u,t}(a) \tilde{N}_{v,t}(a) \sum_k \lambda_k^2 q_k^2(\theta_a).
\]

Furthermore,

\[
\left| \sum_{a \in [n]} \tilde{N}_{u,t}(a) \tilde{N}_{v,t}(a) \sum_k \lambda_k^2 q_k^2(\theta_a) \right| \leq \frac{B^2 \sum_k \lambda_k^2}{\max(|S_u,t|, |S_v,t|)} = O((|S_u,t| |S_v,t|)^{-1/2}).
\]
By combining (7.35), (7.37), and (7.38), it follows that conditioned on \( F(u, v, s, \ell, x, \delta) \), with probability
\[
1 - 2 \exp \left( -n^2 \psi (1 - o(1)) \right) - 2 \exp (n^2 \psi) - O(n^{-6}),
\]
it holds that
\[
\left| \frac{1}{\binom{|S|}{2}} p^2 |S_{u,t}| |S_{v,t}| \sum_{\alpha < \beta \in S \times S} T(\alpha, \beta) - \tilde{N}_{u,t}^T Q^T \Lambda^2 Q \tilde{N}_{v,t} \right|
\leq 2n^\psi \left( \binom{|S|}{2} p^2 |S_{u,t}| |S_{v,t}| \right)^{-1/2} + O(n^{-6}) + o \left( \left( \binom{|S|}{2} p^2 |S_{u,t}| |S_{v,t}| \right)^{-1/2} \right)
+ \frac{\sqrt{S} n^\psi}{|S|^{1/2}} + \frac{B^2 \left( \sum_k \lambda_k^2 \right)}{\max(|S_{u,t}|, |S_{v,t}|)}
\leq O \left( \frac{n^\psi}{\binom{|S|}{2} p^2 |S_{u,t}| |S_{v,t}|} \right)^{1/2} + O \left( \frac{n^\psi}{|S|^{1/2}} \right) + O \left( \frac{1}{\binom{|S|}{2} p^2 |S_{u,t}| |S_{v,t}|} \right)^{1/2}.
\]
Note that the third term is dominated by the first term as \(|S| p = o(1)\). This completes the proof of Lemma 7.7.

\[\square\]

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

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

*Proof.* To compute the variance of \( T(\alpha, \beta) \) conditioned on \( F, \{\theta_i\}_{i \in [n]} \), 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 \((a', b') = (b, a)\),

\[
|\text{Cov}[M_B(a, (\alpha, \beta)), M_B(b, (\alpha, \beta)), M_B(a', (\alpha, \beta)), M_B(b', (\alpha, \beta))]|
= \text{Var}[M_B(a, (\alpha, \beta)) M_B(b, (\alpha, \beta))]
\leq E[M_B^2(a, (\alpha, \beta)) M_B^2(b, (\alpha, \beta))] \leq p^2;
\]

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

\[
|\text{Cov}[M_B(a, (\alpha, \beta)), M_B(b, (\alpha, \beta)), M_B(a', (\alpha, \beta)), M_B(b', (\alpha, \beta)) \mid F, \{\theta_i\}_{i \in [n]}]|
= |\text{Var}[M_B(x, (\alpha, \beta))]| E[M_B(y, (\alpha, \beta))] E[M_B(z, (\alpha, \beta))]| E[M_B(z, (\alpha, \beta))]
\leq |E[M_B^2(x, (\alpha, \beta))]| E[M_B(y, (\alpha, \beta))] E[M_B(z, (\alpha, \beta))]| E[M_B(z, (\alpha, \beta))]
\leq p^3.
\]

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 $a$, $|N_{u,t}(a)| \leq I_{\{a \in S_{u,t}\}}$. We have dropped the conditioning notation due to the length of the expressions.

\[
\begin{align*}
\text{Var}[T(\alpha, \beta) | \mathcal{F}, \{\theta_i\}_{i \in [n]}] &= \sum_{a \neq b \in [n]} (N^2_{u,t}(a)N^2_{v,t}(b) + N_{u,t}(a)N_{v,t}(b)N_{u,t}(b)N_{v,t}(a)) \text{Var}[M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta))] \\
&+ \sum_{a \neq b \in [n]} \sum_{c \notin \{a, b\}} (N^2_{u,t}(a)N_{v,t}(b)N_{u,t}(c) + N^2_{v,t}(a)N_{u,t}(b)N_{u,t}(c) \\
&+ N_{u,t}(a)N_{u,t}(b)N_{v,t}(a)N_{v,t}(c) + N_{u,t}(a)N_{u,t}(c)N_{v,t}(a)N_{v,t}(b)) \times \text{Var}[M_B(a, (\alpha, \beta))] \\
&\times \mathbb{E}[M_B(b, (\alpha, \beta))\mathbb{E}[M_B(c, (\alpha, \beta))]] \\
&\leq p^2 \sum_{a \neq b \in [n]} (I_{\{a \in S_{u,t}, b \in S_{v,t}\}} + I_{\{a, b\} \subset S_{u,t} \cap S_{v,t}}) \\
&+ p^3 \sum_{a \neq b \in [n]} \sum_{c \notin \{a, b\}} \left( I_{\{a \in S_{u,t}, \{b, c\} \subset S_{v,t}\}} + I_{\{a \in S_{u,t}, \{b, c\} \subset S_{u,t}\}} \\
&+ I_{\{a \in S_{u,t} \cap S_{v,t}, b \in S_{v,t}, c \in S_{v,t}\}} + I_{\{a \in S_{u,t} \cap S_{v,t}, c \in S_{u,t}, b \in S_{v,t}\}} \right) \\
&\leq 2p^2 |S_{u,t}| |S_{v,t}| + 2p^3 |S_{u,t}| |S_{v,t}|^2 + 2p^3 |S_{u,t}|^2 |S_{v,t}| \\
&= 2p^2 |S_{u,t}| |S_{v,t}| (1 + o(1)).
\end{align*}
\]

The first term dominates because $p|S_{u,t}| \leq pm = o(1)$ and $p|S_{v,t}| = o(1)$.

**Lemma 7.9.**

\[
P(|T(\alpha, \beta)| \geq z | \mathcal{F}) \leq (|S_{u,t} \cup S_{v,t}|p)^{\sqrt{z}} (1 + o(1)).
\]

As a result,

\[
P(\bigcup_{\alpha, \beta} \{T(\alpha, \beta) \geq z\} | \mathcal{F}) \leq \left( \frac{|S|}{2} \right) (|S_{u,t} \cup S_{v,t}|p)^{\sqrt{z}} (1 + o(1)),
\]

and

\[
\mathbb{E}[I_{|T(\alpha, \beta)| \geq \phi^2} (|T(\alpha, \beta)| - \phi^2) | \mathcal{F}] \leq (1 + o(1)) \frac{2\phi}{\ln(|S_{u,t} \cup S_{v,t}|^{-1}p^{-1})}(|S_{u,t} \cup S_{v,t}|p)^{\phi}
\]

**Proof.** Let us define

\[
Z_u(\alpha, \beta) = \{a \in [n] \text{ s.t. } a \in S_{u,t}, (a, \alpha, \beta) \in \Omega_1\},
\]

\[
Z_v(\alpha, \beta) = \{b \in [n] \text{ s.t. } b \in S_{v,t}, (b, \alpha, \beta) \in \Omega_1\}.
\]

Furthermore, because $|M_B(\cdot, \cdot, \cdot)| \leq 1$ and $\|N_{u,b}\|_{\infty} \leq 1$, for any $a, b \in [n]$, it follows that

\[
|N_{u,t}(a)N_{v,t}(b)M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta))| \leq I_{\{a \in Z_u(\alpha, \beta)\}}I_{\{b \in Z_v(\alpha, \beta)\}},
\]

which implies

\[
|T(\alpha, \beta)| \leq |Z_u(\alpha, \beta)||Z_v(\alpha, \beta)| \leq |Z_u(\alpha, \beta) \cup Z_v(\alpha, \beta)|^2.
\]
Note that $|Z_u(\alpha, \beta) \cup Z_v(\alpha, \beta)| \sim \text{Binomial}(|S_{u,t} \cup S_{v,t}|, p)$. It follows then that
\[
\mathbb{P}(\{T(\alpha, \beta) \geq z \mid \mathcal{F}\}) \leq \mathbb{P}(\{|Z_u(\alpha, \beta) \cup Z_v(\alpha, \beta)|^2 \geq z \mid \mathcal{F}\}) = \mathbb{P}(\{|Z_u(\alpha, \beta) \cup Z_v(\alpha, \beta)| \geq \sqrt{z} \mid \mathcal{F}\}) \leq (1 - p)^{|S_{u,t} \cup S_{v,t}|} \sum_{i=\lceil \sqrt{z} \rceil}^{\infty} \left(\frac{|S_{u,t} \cup S_{v,t}|}{1 - p}\right)^i \sum_{i=0}^\infty \left(\frac{|S_{u,t} \cup S_{v,t}|}{1 - p}\right)^i (1 + o(1)) \leq (|S_{u,t} \cup S_{v,t}|)^{\sqrt{z}} (1 + o(1)) \leq (|S_{u,t} \cup S_{v,t}|)^{\sqrt{z}} (1 + o(1)) \leq (|S_{u,t} \cup S_{v,t}|)^{\sqrt{z}} (1 + o(1)),
\]
where we used the fact that $|S_{u,t} \cup S_{v,t}| = o(1)$.

We use the bound on the tail probabilities to show that
\[
\mathbb{E}[1_{\{T(\alpha, \beta) \geq \phi^2(\|\|T(\alpha, \beta)\| - \phi^2) \mid \mathcal{F}\}] = \int_0^\infty \mathbb{P}(\{T(\alpha, \beta) \geq \phi^2 + z\}) \, dz 
\leq (1 + o(1)) \int_0^\infty (|S_{u,t} \cup S_{v,t}|)^{\sqrt{\phi^2+z}} \, dz 
\leq (1 + o(1)) \int_0^\infty 2y (|S_{u,t} \cup S_{v,t}|)^y \, dy 
= (1 + o(1))(2\phi) \frac{2\phi}{\ln(|S_{u,t} \cup S_{v,t}|^{-1}(1-p)^{-1})} (|S_{u,t} \cup S_{v,t}|)^{\phi}.
\]

\[ \square \]

### 7.5 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,
\[
d(\theta_u, \theta_v) = \|\Lambda^T Q(e_u - e_v)\|^2 = (e_u - e_v)^T Q^T \Lambda^2 Q(e_u - e_v) = e_u^T Q^T \Lambda^2 Q e_u + e_v^T Q^T \Lambda^2 Q e_v - e_u^T Q^T \Lambda^2 Q e_v - e_v^T Q^T \Lambda^2 Q e_u,
\]
and according to (4.3),
\[
\text{dist}(u, v) = \frac{1}{(|S|)^2} |Z_{uu} + Z_{vv} - Z_{uv} - Z_{vu}|
\]
for $S = S_{s,t} = n \setminus (B_{s,t} \cup B_{v,t} \cup [n/2])$ and
\[
Z_{uv} = \frac{1}{(|S_{s,t}|)^2 |S_{s,t}| |S_{v,t}|} \sum_{\alpha, \beta \in S_{s,t} \times S_{u,t}} \tilde{T}_{uv}(\alpha, \beta).
\]
By Lemma 7.1, event $\mathcal{A}_{u,t}^3(\delta)$ holds with probability at least $1 - O(n \exp(-\Theta(n^{2\kappa})))$. By Lemmas 7.2 and 7.6, conditioned on $\mathcal{A}_{u,t}^3(\delta)$, for

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

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

$$
  |\tilde{N}^T_{u,t}Q^T\Lambda^2Q\tilde{N}_{v,t} - e^T_uQ^T\Lambda^{2(t+1)}Qe_v| 
  \leq \frac{n^\psi}{n^\kappa} \left(16B\lambda_{\text{max}}^{2(t-1)} \left( \sum_{k=1}^r 2\lambda_k^{2(t+1)} \right) (1 - \delta)^{-1}(1 + o(1)) \right).
$$

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

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

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

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

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

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

it holds that

$$
  |\text{dist}(u, v) - d(\theta_u, \theta_v)| = O \left( \frac{n^\psi}{n^\kappa} \right).
$$

This completes the proof of Lemma 6.2. \qed

8 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 $\text{dist}$ here, which combined with Lemma 6.2 will result into the proof of Lemma 6.3.

We study the perturbation in the $\text{dist}$ 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 (3.1), $M_1(u, v, w) = F(u, v, w) + \epsilon_{uvw} = F_r(u, v, w) + \epsilon_{uvw} + \epsilon_{uvw}$, 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 $\epsilon_{uvw}$, bounded in absolute value by $\epsilon$. Note that $\epsilon_{uvw}$ can be any arbitrary (or adversarial), unknown deterministic quantity satisfying $|\epsilon_{uvw}| \leq \epsilon$.

Lemma 8.1 provides a bound on the perturbation in the distance estimate, $\text{dist}$, 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 (4.1), $\delta \in (0, \frac{1}{2})$, and

$$
  x = \frac{16\lambda_{\text{max}}^{2t-2}n^\psi}{(1 - \delta)n^\kappa}.
$$
For any $u \in [n]$, recall the event

$$A'_{u,v,t}(x, \delta) = \cap_{k=1}^{r} \left( A_{u,k,t}^4(x, \delta) \cap A_{v,k,t}^4(x, \delta) \right).$$

(8.1)

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 $\varepsilon \geq 0$. Then for any $u, v \in [n]^2$, the distance estimate $\text{dist}(u, v)$ is perturbed by at most $O(t \varepsilon (1 + \varepsilon)^{2t-1} + t^2 \varepsilon^2 (1 + \varepsilon)^{4t-2})$ with probability at least $1 - \exp\left( -\Omega(n^{2\kappa}) \right) - O(n^{-8})$.

Proof. Recall definition of $\text{dist}$ in (4.3):

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

$$Z_{uv} = \frac{1}{|V_B(u, v, t)| p^2 |S_{u,t}| |S_{v,t}|} \sum_{(\alpha, \beta) \in V_B(u, v, t)} T_{uv}(\alpha, \beta),$$

$$V_B(u, v, t) = \{(\alpha, \beta) \in V_B \text{ s.t. } \alpha \notin B_{u,2t} \cup B_{v,2t} \text{ and } \beta \notin B_{u,2t} \cup B_{v,2t}\},$$

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

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 $\varepsilon \geq 0$. This will induce a bound on the changes in $T_{uv}(\cdot, \cdot)$ which will help bound the change in $Z_{uv}$. By assumption (8.1), $A'_{u,v,t}(x, \delta)$ holds. Conditioned on event $A'_{u,v,t}(x, \delta)$, all the information associated with $B_{u,2t}$ and $B_{v,2t}$ is revealed; however, information about $[n] \setminus (B_{u,2t} \cup B_{v,2t})$ is not. Let $F(u, v, t, x, \delta)$ denote all the information revealed such that event $A'_{u,v,t}(x, \delta)$ holds.

Under $A'_{u,v,t}(x, \delta)$, by definition $A^2_{u,t}(\delta)$ and $A^3_{u,t}(\delta)$ holds. This implies that for $S \equiv V_B(u, v, t)$,

$$|S| = \frac{n}{2} - o(n) = \frac{n(1-o(1))}{2}.$$

Furthermore, it implies that $|S_{u,t}|$ and $|S_{v,t}|$ are both greater than or equal to $(1 - \delta)2^{2t-3t-1}n^{2\kappa t}(1 - o(1))$. As shown in (7.27),

$$|S|^2 p^2 |S_{u,t}| |S_{v,t}| = \Omega(n^{2\kappa})$$

(8.2)

results from the choice of $t \geq \frac{1}{2\kappa}$, and the fact that $\delta$ and $t$ are both constants.

For given $\alpha \neq \beta \in V_B(u, v, t)$, $T_{uv}(\alpha, \beta)$ 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 $a \notin S_{u,t}$, $N_{v,t}(b) = 0$ if $b \notin S_{v,t}$. For $a \in S_{u,t}$, $N_{u,t}(a)$ is product of $2t$ terms, each bounded in absolute value by $1$: let $N_{u,t}(a) = \prod_{i=1}^{2t} w_i$ with $|w_i| \leq 1$ for all $i \leq 2t$. Let $\varepsilon_i$ be arbitrary, deterministic quantity added to $w_i$ with $|\varepsilon_i| \leq \varepsilon$ for $i \leq 2t$. Then change in $N_{u,t}(a)$ is bounded as

$$\left| \prod_{i=1}^{2t} w_i - \prod_{i=1}^{2t} (w_i + \varepsilon_i) \right| = \left| \sum_{S \subset [2t]: S \neq \emptyset} \prod_{i \in S} \varepsilon_i \prod_{i \in [2t] \setminus S} w_i \right| \leq \sum_{S \subset [2t]: S \neq \emptyset} \prod_{i \in S} |\varepsilon_i| \prod_{i \in [2t] \setminus S} |w_i|$$

$$\leq \sum_{S \subset [2t]: S \neq \emptyset} |S|^{2t} \varepsilon^s \sum_{i=1}^{2t} \binom{2t}{i} \varepsilon^i$$

$$= \varepsilon \left( \sum_{i=0}^{2t-1} \frac{(2t)!}{(2t-i-1)!(i+1)!} \varepsilon^i \right)$$

$$\leq 2t \varepsilon \left( \sum_{i=0}^{2t-1} \frac{(2t-1)!}{((2t-1)-i)!} \varepsilon^i \right) = 2t \varepsilon \left( \sum_{i=0}^{2t-1} \binom{2t-1}{i} \varepsilon^i \right)$$

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

(8.3)
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 \( |M_B(a, (\alpha, \beta))|, |M_B(b, (\alpha, \beta))| \leq 1 \). Further, \( M_B(a, (\alpha, \beta))M_B(b, (\alpha, \beta)) \neq 0 \) only if \( \mathbb{I}_{((a,\alpha,\beta)\in \Omega)} \mathbb{I}_{((b,\alpha,\beta)\in \Omega)} = 1 \). 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 \( \mathbb{I}_{((a,\alpha,\beta)\in \Omega)} \mathbb{I}_{((b,\alpha,\beta)\in \Omega)} O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2) \). Therefore, we can bound the change in \( Z_{uv} \) by

\[
\frac{O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)}{|S|^2p^2|S_{u,t}||S_{v,t}|} \left( \sum_{a \in S_{u,t}, b \in S_{v,t}, (\alpha, \beta) \in S} \mathbb{I}_{((a,\alpha,\beta)\in \Omega)} \mathbb{I}_{((b,\alpha,\beta)\in \Omega)} \left( 1_{a \neq b} \right) \right)
\]

(8.4)

\[
= \frac{O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2)}{|S|^2p^2|S_{u,t}||S_{v,t}|} \sum_{(\alpha, \beta) \in S} X_{\alpha \beta}
\]

(8.5)

where

\[
X_{\alpha \beta} = \sum_{a \in S_{u,t}, b \in S_{v,t}, a \neq b} \mathbb{I}_{((a,\alpha,\beta)\in \Omega)} \mathbb{I}_{((b,\alpha,\beta)\in \Omega)}
\]

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

\[
\mathbb{P}\left( \bigcup_{\alpha, \beta} \{ X_{\alpha \beta} \geq \phi^2 \} \mid \mathcal{F} \right) \leq \left( \frac{|S|}{2} \right)^2 \left( |S_{u,t} \cup S_{v,t}|p \right)^\phi (1 + o(1)) \quad (8.6)
\]

\[
\mathbb{E}[\mathbb{I}_{X_{\alpha \beta} \geq \phi^2} (X_{\alpha \beta} - \phi^2) \mid \mathcal{F}] \leq (1 + o(1)) \frac{2\phi}{\ln(|S_{u,t} \cup S_{v,t}|^{-1}p^{-1})} (|S_{u,t} \cup S_{v,t}|p)^\phi \quad (8.7)
\]

We define \( \tilde{X}_{\alpha \beta} = \min(X_{\alpha \beta}, \phi^2) \) for \( \phi = [16/(1 - 2\kappa)] \) so that \( \mathbb{E}[\tilde{X}_{\alpha \beta}] \leq \phi^2 \). By (8.6), and the choice of \( \phi \) along with the conditions from \( \mathcal{F} \) that guarantee \( |S| = \Theta(n) \) and naively \( |S_{u,t} \cup S_{v,t}| = O(n) \),

\[
\mathbb{P}\left( \sum_{\alpha, \beta} X_{\alpha \beta} \neq \sum_{\alpha, \beta} \tilde{X}_{\alpha \beta} \mid \mathcal{F} \right) \leq \left( \frac{|S|}{2} \right)^2 \left( |S_{u,t} \cup S_{v,t}|p \right)^\phi (1 + o(1)) = O(n^{-6}). \quad (8.8)
\]

By (8.7),

\[
|\mathbb{E}[X_{\alpha \beta} \mid \mathcal{F}] - \mathbb{E}[\tilde{X}_{\alpha \beta} \mid \mathcal{F}]| \leq \mathbb{E}[\mathbb{I}_{X_{\alpha \beta} \geq \phi^2} (X_{\alpha \beta} - \phi^2) \mid \mathcal{F}] \leq (1 + o(1)) \frac{2\phi}{\ln(|S_{u,t} \cup S_{v,t}|^{-1}p^{-1})} (|S_{u,t} \cup S_{v,t}|p)^\phi = O(n^{-6}).
\]

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

\[
\text{Var}[\tilde{X}_{\alpha \beta} \mid \mathcal{F}, \{\theta_i\}_{i \in [n]}] \leq \text{Var}[X_{\alpha \beta} \mid \mathcal{F}, \{\theta_i\}_{i \in [n]}] \leq 2p^2|S_{u,t}||S_{v,t}|(1 + o(1)).
\]

By Bernstein’s inequality, for \( z = 2n^\psi |S|p(|S_{u,t}||S_{v,t}|)^{1/2} \) for \( \psi = 0(\kappa), \)

\[
\mathbb{P}\left( \sum_{\alpha, \beta \in S} (\tilde{X}_{\alpha \beta} - \mathbb{E}[\tilde{X}_{\alpha \beta}]) > z \right) \leq \exp \left( -\frac{3z^2}{2z\phi^2 + 12(1 + o(1))|S|^2p^2|S_{u,t}||S_{v,t}|} \right)
\]

(8.9)

\[
= \exp(-n^2\psi(1 - o(1))). \quad (8.10)
\]
By (8.5), (8.8), (8.9), and (8.10), given \( A'_{u,v,t}(x, \delta) \) holds, with probability \( 1 - \exp \left( -n^{2\kappa}(1 - o(1)) \right) - O(n^{-6}) \), the change in \( Z_{uv} \) is bounded above by

\[
O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2) \sum_{\alpha, \beta \in S} X_{\alpha\beta} = O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2) \sum_{\alpha, \beta \in S} \tilde{X}_{\alpha\beta} = O(\Delta(t, \varepsilon) + \Delta(t, \varepsilon)^2) \left( (\mathbb{E} \sum_{\alpha, \beta \in S} X_{\alpha\beta} \mid \mathcal{F}) + 2n^\psi |S|^p(|S_{u,t}| ||S_{v,t}|)^{1/2} \right).
\]

By (8.2), this choice of \( 2n^\psi |S|^p(|S_{u,t}| ||S_{v,t}|)^{1/2} = o(|S|^2p^2|S_{u,t}| ||S_{v,t}|) \) for \( \psi \in (0, \kappa) \). Finally, we use the bound that \( \mathbb{E} |\sum_{\alpha, \beta \in S} X_{\alpha\beta} \mid \mathcal{F} = (\frac{|S|^p}{2})^2|S_{u,t}| ||S_{v,t}| \) to argue that with high probability the change in \( Z_{uv} \) is bounded above by

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

This completes the proof of Lemma 8.1.

\[
\square
\]

### 8.1 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 \( x, \delta \) as considered in statement of Lemma 8.1, holds with probability at least \( 1 - 4\tau \exp(-n^{2\psi}(1 - o(1))) \). And \( \text{dist} \) (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 \( \text{dist} \) is further perturbed by \( O \left( t^2(1 + \varepsilon)^{2t-1} + \varepsilon^2(1 + \varepsilon)^{4t-2} \right) \) with probability at least \( 1 - O \left( \exp(-n^{2\psi}(1 - o(1))) \right) - O \left( n^{-6}\right) \).

Putting these together, we conclude the claim of Lemma 6.3.

### Acknowledgements

We gratefully acknowledge funding from the NSF under grants CCF-1948256 and CNS-1955997.

### References

[1] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In Conference on Learning Theory, 2016.

[2] Srinadh Bhojanapalli and Sujay Sanghavi. A new sampling technique for tensors. arXiv preprint arXiv:1502.05023, 2015.

[3] Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In Advances in Neural Information Processing Systems, pages 4715–4726, 2017. Also available at ArXiv.

[4] Christian Borgs, Jennifer Chayes, Devavrat Shah, and Christina Lee Yu. Iterative collaborative filtering for sparse matrix estimation. Operations Research, forthcoming 2021.
[5] Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6), 2009.

[6] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[7] Emmanuel J Candes and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[8] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[9] Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[10] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[11] Shmuel Friedland and Lek-Heng Lim. Nuclear norm of higher-order tensors. *arXiv preprint arXiv:1410.6072*, 2014.

[12] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[13] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[14] David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 1992.

[15] Moritz Hardt. Understanding alternating minimization for matrix completion. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE, 2014.

[16] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*, 58:13–30, 1963.

[17] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[18] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, 2014.

[19] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[20] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[21] Y. Li, D. Shah, D. Song, and C. L. Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, March 2020.
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 $E[M_2(a, b, c)] = f(\theta_a, \theta_b, \theta_c)$. Conditioned on $\Omega_2$, by definition of $\hat{F}$ and by
By definition, assuming properties 6.1 and 6.2, it follows that

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

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) - d(u, a)| \leq \Delta and \hat{d}(u, a) \leq \eta \implies d(u, a) \leq \eta + \Delta, similarly \delta(v, b), d(w, c) \leq \eta + \Delta. As per (3.1), we have that \text{Var}[M_2(a, b, c)] \leq \sigma^2 for all (a, b, c) \in \Omega_2. Define \( \nu_{uvw} = \{(a, b, c) \in [n]^3 : d(u, a) < \eta - \Delta, d(v, b) < \eta - \Delta, d(w, c) < \eta - \Delta\} \). Assuming property 6.3,

\[
|\nu_{uvw}| = |\{a \in [n] : d(u, a) < \eta - \Delta\}| \cdot |\{b \in [n] : d(v, b) < \eta - \Delta\}| \cdot |\{c \in [n] : d(w, c) < \eta - \Delta\}| \\
\geq (\text{meas}(\eta - \Delta)n)^3.
\]

By the Bernoulli sampling model, each tuple \((a, b, c) \in [n]^3\) belongs to \(\Omega_2\) with probability \(p\) independently. By a straightforward application of Chernoff’s bound, it follows that for any \(\delta \in (0, 1),\)

\[
\mathbb{P}\left(|\Omega_2 \cap \nu_{uvw}| \leq (1 - \delta) (\text{meas}(\eta - \Delta)n)^3\right) \leq \exp\left(-\frac{\delta^2 p (\text{meas}(\eta - \Delta)n)^3}{2}\right).
\]

(11)

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 \(1 - \exp\left(-\frac{\delta^2 p (\text{meas}(\eta - \Delta)n)^3}{2}\right),\)

\[
|\Omega_{uvw}| = |\{(a, b, c) \in \Omega_2 : d(u, a) < \eta, d(v, b) < \eta, d(w, c) < \eta\}| \\
\geq |\{(a, b, c) \in \Omega_2 : d(u, a) < \eta - \Delta, d(v, b) < \eta - \Delta, d(w, c) < \eta - \Delta\}| \\
= |\Omega_2 \cap \nu_{uvw}| \\
\geq (1 - \delta)p (\text{meas}(\eta - \Delta)n)^3.
\]

Define the event \(H = |\Omega_{uvw}| \geq (1 - \delta)p (\text{meas}(\eta - \Delta)n)^3\). It follows that \(\mathbb{P}(H^c) \leq \exp\left(-\frac{\delta^2 p (\text{meas}(\eta - \Delta)n)^3}{2}\right)\)

By definition, \(F(u, v, w) = f(\theta_u, \theta_v, \theta_w) \in [0, 1]\) for all \(u, v, w \in [n]\). Therefore,

\[
\mathbb{E}[(\hat{F}(u, v, w) - f(\theta_u, \theta_v, \theta_w))^2] \\
\leq \mathbb{E}[(\hat{F}(u, v, w) - f(\theta_u, \theta_v, \theta_w))^2 \mid H] + \mathbb{P}(H^c) \\
\leq \text{bias}^2(\eta + \Delta) + \frac{1}{(1 - \delta)p (\text{meas}(\eta - \Delta)n)^3} + \exp\left(-\frac{\delta^2 p (\text{meas}(\eta - \Delta)n)^3}{2}\right).
\]

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 \(\alpha_2\) 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 $\Omega_1$. 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

$$\mathbb{P}\left(\frac{\left|\sum_{(a, b, c) \in \Omega_2^{uvw}} (M(a, b, c) - F(a, b, c))\right|}{|\Omega_2^{uvw}|} \geq \delta' \mid \mathcal{H}\right) \leq \exp\left(-\delta'^2 (1 - \delta) p (\text{meas}(\eta - \Delta) n)^3\right).$$

Therefore,

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

with probability at least

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

This completes the proof of Lemma 6.1.

**Lemma 2.** 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 $s$,

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

(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 $\exp(-a\rho) \leq \frac{1}{2}$,

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

$$\leq \exp(-a\rho) \sum_{s=0}^{r-1} \exp(-aps) \leq \exp(-a\rho) \sum_{s=0}^{r-1} 2^{-s} \leq 2\exp(-a\rho).$$

\[\square\]
Lemma 3. If $P(|X| \geq z) \leq c \exp(-\frac{z^2}{Q})$, then for all $\lambda \in \mathbb{R}$,

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

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

Proof.

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

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

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

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

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

Lemma 4. 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

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

(.12)

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

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

(.13)

The proof follows directly from an implication of Azuma-Hoeffding’s inequality. For example, see [32, Example 2.23] for a proof.