Provable Tensor Factorization with Missing Data

Prateek Jain[*]and Sewoong Oh[†]

Abstract

We study the problem of low-rank tensor factorization in the presence of missing data. We ask the following question: how many sampled entries do we need, to efficiently and exactly reconstruct a tensor with a low-rank orthogonal decomposition? We propose a novel alternating minimization based method which iteratively refines estimates of the singular vectors. We show that under certain standard assumptions, our method can recover a three-mode $n \times n \times n$ dimensional rank-r tensor exactly from $O(n^{3/2} r^5 \log^4 n)$ randomly sampled entries. In the process of proving this result, we solve two challenging sub-problems for tensors with missing data. First, in the process of analyzing the initialization step, we prove a generalization of a celebrated result by Szemerédie et al. on the spectrum of random graphs. Next, we prove global convergence of alternating minimization with a good initialization. Simulations suggest that the dependence of the sample size on dimensionality n is indeed tight.

1 Introduction

Several real-world applications routinely encounter multi-way data with structure which can be modeled as low-rank tensors. Moreover, in several settings, many of the entries of the tensor are missing, which motivated us to study the problem of low-rank tensor factorization with missing entries. For example, when recording electrical activities of the brain, the electroencephalography (EEG) signal can be represented as a three-way array (temporal, spectral, and spatial axis). Oftentimes signals are lost due to mechanical failure or loose connection. Given numerous motivating applications, several methods have been proposed for this tensor completion problem. However, with the exception of 2-way tensors (i.e., matrices), the existing methods for higher-order tensors do not have theoretical guarantees and typically suffer from the curse of local minima.

In general, finding a factorization of a tensor is an NP-hard problem, even when all the entries are available. However, it was recently discovered that by restricting attention to a sub-class of tensors such as low-CP rank orthogonal tensors [1] or low-CP rank incoherent[1] tensors [2], one can efficiently find a provably approximate factorization. In particular, exact recovery of the factorization is possible for a tensor with a low-rank orthogonal CP decomposition [1]. We ask the question of recovering such a CP-decomposition when only a small number of entries are revealed, and show that exact reconstruction is possible even when we do not observe any entry in most of the fibers.

Problem formulation. We study tensors that have an orthonormal CANDECOMP/PARAFAC (CP) tensor decomposition with a small number of components. Moreover, for simplicity of notation and exposition, we only consider symmetric third order tensors. We would like to stress that our techniques generalizes easily to handle *non-symmetric tensors* as well as *higher-order* tensors. Formally, we assume that the true

---

[*]Microsoft Research, `prajain@microsoft.com`

[†]Department of Industrial and Enterprise Systems Engineering, Univrsity of Illinois at Urbana-Champaign, `swoh@illinois.edu`

[1]The notion of incoherence we assume in (2) can be thought of as incoherence between the fibers and the standard basis vectors.

tensor T has the the following form:

$$T = \sum_{\ell=1}^{r} \sigma_\ell(\mathbf{u}_\ell \otimes \mathbf{u}_\ell \otimes \mathbf{u}_\ell) \in \mathbb{R}^{n \times n \times n} , \tag{1}$$

with $r \ll n$, $u_\ell \in \mathbb{R}^n$ with $\|u_\ell\| = 1$, and u_ℓ's are orthogonal to each other. We let $U \in \mathbb{R}^{n \times r}$ be a tall-orthogonal matrix where u_ℓ's is the ℓ-th column of U and $U_i \perp U_j$ for $i \neq j$. We use $\otimes$ to denote the standard outer product such that the (i, j, k)-th element of T is given by: $T_{ijk} = \sum_a \sigma_a U_{ia} U_{ja} U_{ka}$. We further assume that the u_i's are unstructured, which is formalized by the notion of *incoherence* commonly assumed in matrix completion problems. The *incoherence* of a symmetric tensor with orthogonal decomposition is

$$\mu(T) \equiv \max_{i \in [n], \ell \in [r]} \sqrt{n} \, |U_{i\ell}| , \tag{2}$$

where $[n] = \{1, \ldots, n\}$ is the set of the first n integers. Tensor completion becomes increasingly difficult for tensors with larger $\mu(T)$, because the 'mass' of the tensor can be concentrated on a few entries that might not be revealed. Out of n^3 entries of T, a subset $\Omega \subseteq [n] \times [n] \times [n]$ is revealed. We use $\mathcal{P}_\Omega(\cdot)$ to denote the projection of a matrix onto the revealed set such that

$$\mathcal{P}_\Omega(T)_{ijk} = \begin{cases} T_{ijk} & \text{if } (i, j, k) \in \Omega , \\ 0 & \text{otherwise} . \end{cases}$$

We want to recover T exactly using the given entries $(P_\Omega(T))$. We assume that each (i, j, k) for all $i \leq j \leq k$ is included in Ω with a *fixed probability* p (since T is symmetric, we include all permutations of (i, j, k)). This is equivalent to fixing the total number of samples $|\Omega|$ and selecting Ω *uniformly at random* over all $\binom{n^3}{|\Omega|}$ choices. The goal is to ensure exact recovery with high probability and for $|\Omega|$ that is sub-linear in the number of entries (n^3).

Notations. For a tensor $T \in \mathbb{R}^{n \times n \times n}$, we define a linear mapping using $U \in \mathbb{R}^{n \times m}$ as $T[U, U, U] \in \mathbb{R}^{m \times m \times m}$ such that $T[U, U, U]_{ijk} = \sum_{a,b,c} T_{abc} U_{ai} U_{bj} U_{ck}$. The operator norm of a tensor is $\|T\|_2 = \max_{\|x\|=1} T[x, x, x]$. The Frobenius norm of a tensor is $\|T\|_F = (\sum_{i,j,k} T_{ijk}^2)^{1/2}$. The Euclidean norm of a vector is $\|\mathbf{u}\|_2 = (\sum_i \mathbf{u}_i^2)^{1/2}$. We use C, C' to denote any positive numerical constants and the actual value might change from line to line.

1.1 Algorithm

Ideally, one would like to minimize the rank of a tensor that explains all the sampled entries.

$$\underset{\widehat{T}}{\text{minimize}} \quad \text{rank}(\widehat{T}) \tag{3}$$

$$\text{subject to} \quad T_{ijk} = \widehat{T}_{ijk} \text{ for all } (i, j, k) \in \Omega .$$

However, even computing the rank of a tensor is NP-hard in general, where the rank is defined as the minimum r for which CP-decomposition exists [3]. Instead, we fix the rank of $\widehat{T}$ by explicitly modeling $\widehat{T}$ as $\widehat{T} = \sum_{\ell \in [r]} \sigma_\ell(\mathbf{u}_\ell \otimes \mathbf{u}_\ell \otimes \mathbf{u}_\ell)$, and solve the following problem:

$$\underset{\widehat{T}, \text{rank}(\widehat{T})=r}{\text{minimize}} \left\| \mathcal{P}_\Omega(T) - \mathcal{P}_\Omega(\widehat{T}) \right\|_F^2 = \underset{\{\sigma_\ell, \mathbf{u}_\ell\}_{\ell \in [r]}}{\text{minimize}} \left\| \mathcal{P}_\Omega(T) - \mathcal{P}_\Omega\Big(\sum_{\ell \in [r]} \sigma_\ell(\mathbf{u}_\ell \otimes \mathbf{u}_\ell \otimes \mathbf{u}_\ell) \Big) \right\|_F^2 \tag{4}$$

Recently, [4, 5] showed that an alternating minimization technique, can recover a matrix with missing entries exactly. We generalize and modify the algorithm for the case of higher order tensors and study it rigorously for tensor completion. However, due to special structure in higher-order tensors, our algorithm as well as analysis is significantly different than the matrix case (see Section 2.2 for more details).

To perform the minimization, we repeat the outer-loop getting refined estimates for all r components. In the inner-loop, we loop over each component and solve for $\mathbf{u}_q$ while fixing the others $\{\mathbf{u}_\ell\}_{\ell \neq q}$. More precisely, we set $\widehat{T} = \mathbf{u}_q^{t+1} \otimes \mathbf{u}_q \otimes \mathbf{u}_q + \sum_{\ell \neq q} \sigma_\ell \mathbf{u}_\ell \otimes \mathbf{u}_\ell \otimes \mathbf{u}_\ell$ in (4) and then find optimal $\mathbf{u}_q^{t+1}$ by minimizing the least squares objective given by (4). That is, each inner iteration is a simple least squares problem over the known entries, hence can be implemented efficiently and is also embarrassingly parallel.

Algorithm 1 Alternating Minimization for Tensor Completion

1: Input: $P_\Omega(T)$, Ω, r, τ
2: Initialize with $[(\mathbf{u}_1^0, \sigma_1), (\mathbf{u}_2^0, \sigma_2), \ldots, (\mathbf{u}_r^0, \sigma_r)] = RTPM(P_\Omega(T), r)$ (RTPM of [1])
3: $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] = \text{Threshold}([\mathbf{u}_1^0, \mathbf{u}_2^0, \ldots, \mathbf{u}_r^0])$ (Clipping scheme of [4])
4: **for all** $t = 1, 2, \ldots, \tau$ **do**
5:     /*OUTER LOOP */
6:     **for all** $q = 1, 2, \ldots, r$ **do**
7:        /*INNER LOOP*/
8:        $\hat{\mathbf{u}}_1^{t+1} = \arg\min_{\mathbf{u}_q^{t+1}} \|\mathcal{P}_\Omega(T - \mathbf{u}_q^{t+1} \otimes \mathbf{u}_q \otimes \mathbf{u}_q - \sum_{\ell \neq q} \sigma_\ell \cdot \mathbf{u}_\ell \otimes \mathbf{u}_\ell \otimes \mathbf{u}_\ell)\|_F^2$
9:        $\sigma_q^{t+1} = \|\hat{\mathbf{u}}_q^{t+1}\|_2$
10:        $\mathbf{u}_q^{t+1} = \hat{\mathbf{u}}_1^{t+1} / \|\hat{\mathbf{u}}_1^{t+1}\|_2$
11:     **end for**
12:     $[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r] \leftarrow [\mathbf{u}_1^{t+1}, \mathbf{u}_2^{t+1}, \ldots, \mathbf{u}_r^{t+1}]$
13:     $[\sigma_1, \sigma_2, \ldots, \sigma_r] \leftarrow [\sigma_1^{t+1}, \sigma_2^{t+1}, \ldots, \sigma_r^{t+1}]$
14: **end for**
15: Output: $\widehat{T} = \sum_{q \in [r]} \sigma_q(\mathbf{u}_q \otimes \mathbf{u}_q \otimes \mathbf{u}_q)$

The *main novelty* in our approach is that we refine all r components iteratively as opposed to the sequential deflation technique used by the existing methods for tensor decomposition (for fully observed tensors). In sequential deflation methods, components $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ are estimated sequentially and estimate of say $\mathbf{u}_2$ is not used to refine $\mathbf{u}_1$. In contrast, our algorithm iterates over all r estimates in the inner loop, so as to obtain refined estimates for all $\mathbf{u}_i$'s in the outer loop. We believe that such a technique could be applied to improve the error bounds of (fully observed) tensor decomposition methods as well.

As our method is directly solving a non-convex problem, it can easily get stuck in local minima. The key reason our approach can overcome the curse of local minima is that we start with a provably good initial point which is only a small distance away from the optima. To obtain such an initial estimate, we compute a low-rank approximation of the observed tensor using *Robust Tensor Power Method* (RTPM) [1]. RTPM is a generalization of the widely used power method for computing leading singular vectors of a matrix and can approximate the largest singular vectors up to the spectral norm of the "error" tensor. Hence, the challenge is to show that the error tensor has small spectral norm (see Theorem 2.1). We perform a thresholding step similar to [4] (see Lemma A.4) after the RTPM step to ensure that the estimates we get are incoherent, which is critical for our analysis.

Our analysis requires the sampled entries Ω to be independent of the current iterates $\mathbf{u}_i, \forall i$, which in general is not possible as $\mathbf{u}_i$'s are computed using Ω. To avoid this issue, we divide the given samples (Ω) into equal $r \cdot \tau$ parts randomly where τ is the number of outer loops (see Algorithm 1).

1.2   Main Result

Theorem 1.1. *Consider any rank-r symmetric tensor $T \in \mathbb{R}^{n \times n \times n}$ with an orthogonal CP decomposition in (1) satisfying μ-incoherence as defined in (2). For any positive $\varepsilon > 0$, there exists a positive numerical*

constant C such that if entries are revealed with probability

$$p \;\geq\; C\,\frac{\mu^6\,r^5\,\sigma_{\max}^4\,(\log n)^4\,\log(r\|T\|_F/\varepsilon)}{\sigma_{\min}^4\,n^{3/2}}\;,$$

then the following holds with probability at least $1 - n^{-5}\log_2(4\sqrt{r}\,\|T\|_F/\varepsilon)$:

- *the problem* (3) *has a unique optimal solution; and*

- $\log_2\!\left(\frac{4\sqrt{r}\,\|T\|_F}{\varepsilon}\right)$ *iterations of Algorithm 1 produces an estimate $\widehat{T}$ s.t. $\|T - \widehat{T}\|_F \leq \varepsilon$.*

Note that the above result can be generalized to k-mode tensors in a straightforward manner, where exact recovery is guaranteed if, $p \geq C\,\frac{\mu^6\,r^5\,\sigma_{\max}^4\,(\log n)^4\,\log(r\|T\|_F/\varepsilon)}{\sigma_{\min}^4\,n^{k/2}}$. However, for simplicity of notation and to emphasize key points of our proof we present our proof for 3-mode tensors only in Section 2.3.

We provide a proof of Theorem 1.1 in Section 2. For an incoherent, well-conditioned, and low-rank tensor with $\mu = O(1)$ and $\sigma_{\min} = \Theta(\sigma_{\max})$, alternating minimization requires $O(r^5 n^{3/2}(\log n)^4)$ samples to get within an arbitrarily small normalized error. This is a vanishing fraction of the total number of entries n^3. Each step in the alternating minimization requires $O(r|\Omega|)$ operations, hence the alternating minimization only requires $O(r|\Omega|\log(r\|T\|_F/\varepsilon))$ operations. The initialization step requires $O(r^c|\Omega|)$ operations for some positive constant numerical c. When $r \ll n$, the computational complexity scales linearly in the sample size up to a logarithmic factor.

A fiber in a third order tensor is an n-dimensional vector defined by fixing two of the axes and indexing over remaining one axis. The above theorem implies that among n^2 fibers of the form $\{T[\mathbb{I}, e_j, e_k]\}_{j,k \in [n]}$, it is sufficient to have only $O(n^{3/2}(\log n)^4)$ fibers with *any* samples. Most of the fibers are not sampled at all and, perhaps surprisingly, our approach can still recover the original low-rank tensor. This should be compared to the matrix completion setting where all fibers are required to have at least one sample.

However, unlike matrices, the fundamental limit of higher order tensor completion is not known. Building on the percolation of Erdös-Renýi graphs and the coupon-collectors problem, it is known that matrix completion has multiple rank-r solutions when the sample size is less than $C\mu rn \log n$ [6], hence exact recovery is impossible. But, such arguments do not generalize directly to higher order; see Section 2.5 for more discussion. Interestingly, simulations in Section 1.3 suggests that for $r = O(\sqrt{n})$, the sample complexity scales as $(r^{1/2}n^{3/2}\log n)$. That is, assuming the sample complexity provided by simulations is correct, our result achieves optimal dependence on n (up to log factors). However, the dependency on r is sub-optimal (see Section 2.5 for a discussion).

1.3 Empirical Results

Theorem 1.1 guarantees exact recovery when $p \geq Cr^5(\log n)^4/n^{3/2}$. Numerical experiments show that the average recovery rate converges to a universal curve over α, where $p^* = \alpha r^{1/2} \ln n / ((1 - \rho)n^{3/2})$ in Figure 1. Our bound is tight in its dependency n up to a poly-logarithmic factor, but is loose in its dependency in the rank r. Further, it is able to recover the original matrix exactly even when the factors are not strictly orthogonal.

We generate orthogonal matrices $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r] \in \mathbb{R}^{n \times r}$ uniformly at random with $n = 50$ and $r = 3$ unless specified otherwise. For a rank-r tensor $T = \sum_{i=1}^{r} \mathbf{u}_i \otimes \mathbf{u}_i \otimes \mathbf{u}_i$, we randomly reveal each entry with probability p. A tensor is exactly recovered if the normalized root mean squared error, RMSE $= \|T - \hat{T}\|_F/\|T\|_F$, is less than 10^{-7}. Varying n and r, we plot the recovery rate averaged over 100 instances as a function of α. The degrees of freedom in representing a symmetric tensor is $rn - r^2$. Hence for large, r we need number of samples scaling as r. Hence, the current dependence of $p^* = O(\sqrt{r})$ can only hold for $r = O(n)$. For not strictly orthogonal factors, the algorithm is robust. A more robust approach for finding an initial guess could improve the performance significantly, especially for non-orthogonal tensors.
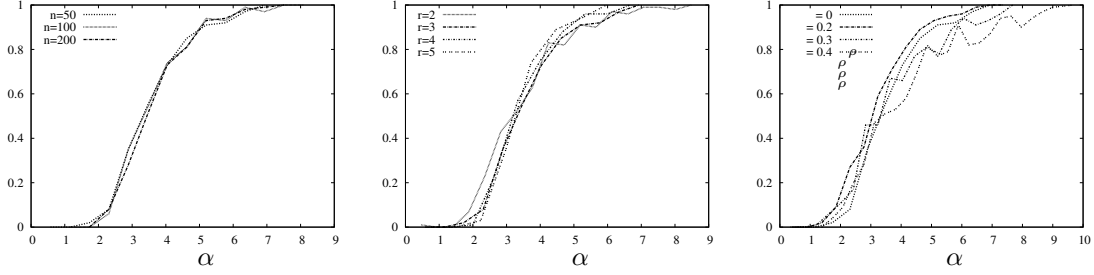
Figure 1: Average recovery rate converges to a universal curve over α when $p = \alpha r^{1/2} \ln n / ((1-\rho) n^{3/2})$, where $\rho = \max_{i \neq j \in [r]} \langle \mathbf{u}_i, \mathbf{u}_j \rangle$ and $r = O(\sqrt{n})$.

1.4 Related Work

Tensor decomposition and completion: The CP model proposed in [7, 8, 9] is a multidimensional generalization of singular value decomposition of matrices. Computing the CP decomposition involves two steps: first apply a whitening operator to the tensor to get a lower dimensional tensor with orthogonal CP decomposition. Such a whitening operator only exists when $r \leq n$. Then, apply known power-method techniques for exact orthogonal CP decomposition [1]. We use this algorithm as well as the analysis for the initial step of our algorithm. For motivation and examples of orthogonal CP models we refer to [10, 1].

Recently, many heuristics for tensor completion have been developed such as the weighted least squares [11], Gauss-Newton [12], alternating least-squares [13, 14], trace norm minimization [15]. However, to the best of our knowledge, there is no tensor completion method with provable guarantees. In a different context, [16] show that minimizing a weighted trace norm of flattened tensor provides exact recovery using $O(rn^{3/2})$ samples, but each observation needs to be a dense random projection of the tensor as opposed to observing just a single entry, which is the case in the tensor completion problem.

Relation to matrix completion: Matrix completion has been studied extensively in the last decade since the seminal paper by Candes and Recht [17]. Since then, several provable approaches have been developed, such as, nuclear norm minimization [17], OptSpace [18], and Alternating Minimization [4]. However, several aspects of tensor factorization makes it challenging to adopt matrix completion algorithms and analysis techniques directly.

First, there is no natural convex surrogate of the tensor rank function and developing such a function is in fact a topic of active research [19, 16] Next, even when all entries are revealed, tensor decomposition methods such as simultaneous power iteration are known to get stuck at local extrema, making it challenging to apply matrix decomposition methods directly. Third, for the initialization step, the best low-rank approximation of a matrix is unique and finding it is trivial. However, for tensors, finding the best low-rank approximation is notoriously difficult.

On the other hand, some aspects of tensor decomposition makes it possible to prove stronger results. Matrix completion aims to recover the underlying matrix only, since the factors are not uniquely defined due to invariance under rotations. However, for orthogonal CP models, we can hope to recover the individual singular vectors $\mathbf{u}_i$'s exactly. In fact, Theorem 1.1 shows that our method indeed recovers the individual singular vectors exactly.

Spectral analysis of tensors and hypergraphs: Theorem 2.1 and Lemma 2.2 should be compared to copious line of work on spectral analysis of matrices [20, 18], with an important motivation of developing fast algorithms for low-rank matrix approximations. We prove an analogous guarantee for higher order tensors and provide a fast algorithm for low-rank tensor approximation. Theorem 2.1 is also a generalization of the celebrated result of Friedman-Kahn-Szemerédi [21] and Feige-Ofek [22] on the second eigenvalue of random graphs. We

5

provide an upper bound the largest second eigenvalue of a random hypergraph, where each edge includes three nodes and each of the $\binom{n}{3}$ edges is selected with probability p.

2 Analysis of the Alternating Minimization Algorithm

In this section, we provide a proof of Theorem 1.1 and the proof sketches of the required main technical theorems. Formal proofs of the technical theorems and lemmas are provided in the appendix. There are two key components: a) the analysis of the initialization step (Section 2.1); and b) the convergence of alternating minimization given a sufficiently accurate initialization (Section 2.2). We use these two analyses to prove Theorem 1.1 in Section 2.3.

2.1 Initialization Analysis

We first show that $(1/p)\mathcal{P}_\Omega(T)$ is close to T in spectral norm, and use it bound the error of robust power method applied directly to $\mathcal{P}_\Omega(T)$. The normalization by $(1/p)$ compensates for the fact that many entries are missing. For a proof of this theorem, we refer to Appendix A.

Theorem 2.1 (Initialization). *For $p = \alpha/n^{3/2}$ satisfying $\alpha \geq \log n$, there exists a positive constant $C > 0$ such that, with probability at least $1 - n^{-5}$,*

$$\frac{1}{T_{\max} n^{3/2} p} \|\mathcal{P}_\Omega(T) - p\,T\|_2 \quad \leq \quad \frac{C\,(\log n)^2}{\sqrt{\alpha}} \;, \tag{5}$$

where $T_{\max} \equiv \max_{i,j,k} T_{ijk}$, and $\|T\|_2 \equiv \max_{\|u\|=1} T[u,u,u]$ is the operator norm.

Notice that $T_{\max}$ is the maximum entry in the tensor T and the factor $1/(T_{\max} n^{3/2} p)$ corresponds to normalization with the worst case operator norm of $p\,T$, since $\|pT\|_2 \leq T_{\max} n^{3/2} p$ and the maximum is achieved by $T = T_{\max}(\mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1})$. The following theorem guarantees that $O(n^{3/2}(\log n)^2)$ samples are sufficient to ensure that we get arbitrarily small error. A formal proof is provided in Appendix.

Together with an analysis of *robust tensor power method* [1, Theorem 5.1], the next error bound follows from directly substituting (5) and using the fact that for incoherent tensors $T_{\max} \leq \sigma_{\max}\mu(T)^3 r/n^{3/2}$. Notice that the estimates can be computed efficiently, requiring only $O(\log r + \log\log\alpha)$ iterations, each iteration requiring $O(\alpha n^{3/2})$ operations. This is close to the time required to read the $|\Omega| \simeq \alpha n^{3/2}$ samples. One caveat is that we need to run robust power method poly$(r\log n)$ times, each with fresh random initializations.

Lemma 2.2. *For a μ-incoherent tensor with orthogonal decomposition $T = \sum_{\ell=1}^r \sigma_\ell^*(\mathbf{u}_\ell^* \otimes \mathbf{u}_\ell^* \otimes \mathbf{u}_\ell^*) \in \mathbb{R}^{n \times n \times n}$, there exists positive numerical constants C, C' such that when $\alpha \geq C(\sigma_{\max}/\sigma_{\min})^2 r^5 \mu^6 (\log n)^4$, running $C'(\log r + \log\log\alpha)$ iterations of the robust tensor power method applied to $\mathcal{P}_\Omega(T)$ achieves*

$$\|\mathbf{u}_\ell^* - \mathbf{u}_\ell^0\|_2 \quad \leq \quad C' \frac{\sigma_{\max}^*}{|\sigma_\ell^*|} \frac{\mu^3 r (\log n)^2}{\sqrt{\alpha}} \;,$$

$$\frac{|\sigma_\ell^* - \sigma_\ell|}{|\sigma_\ell^*|} \quad \leq \quad C' \frac{\sigma_{\max}^*}{|\sigma_\ell^*|} \frac{\mu^3 r (\log n)^2}{\sqrt{\alpha}} \;,$$

for all $\ell \in [r]$ with probability at least $1 - n^{-5}$, where $\sigma_{\max}^* = \max_{\ell \in [r]} |\sigma_\ell^*|$ and $\sigma_{\min}^* = \min_{\ell \in [r]} |\sigma_\ell^*|$.

2.2 Alternating Minimization Analysis

We now provide convergence analysis for the alternating minimization part of Algorithm 1 to recover rank-r tensor T. Our analysis assumes that $\|\mathbf{u}_i - \mathbf{u}_i^*\|_2 \leq c\sigma_{\min}/r\sigma_{\max}, \forall i$ where c is a small constant (dependent on

r and the condition number of T). The above mentioned assumption can be satisfied using our initialization analysis and by assuming Ω is large-enough.

At a high-level, our analysis shows that each step of Algorithm 1 ensures geometric decay of a distance function (specified below) which is "similar" to $\max_j \|\mathbf{u}_j^t - \mathbf{u}_j^*\|_2$.

Formally, let $T = \sum_{\ell=1}^r \sigma_\ell^* \cdot \mathbf{u}_\ell^* \otimes \mathbf{u}_\ell^* \otimes \mathbf{u}_\ell^*$. WLOG, we can assume that that $\sigma_\ell^* \leq 1$. Also, let $[U, \ \Sigma] = \{(\mathbf{u}_\ell, \sigma_\ell), 1 \leq \ell \leq r\}$, be the t-th step iterates of Algorithm 1. We assume that $\mathbf{u}_\ell^*, \forall \ell$ are μ-incoherent and $\mathbf{u}_\ell, \forall \ell$ are 2μ-incoherent. Define, $\Delta_\ell^\sigma = \frac{|\sigma_\ell - \sigma_\ell^*|}{\sigma_\ell^*}$, $\mathbf{u}_\ell = \mathbf{u}_\ell^* + \mathbf{d}_\ell$, $(\Delta_\ell^\sigma)^{t+1} = \frac{|\sigma_\ell^{t+1} - \sigma_\ell^*|}{\sigma_\ell^*}$, and $\mathbf{u}_\ell^{t+1} = \mathbf{u}_\ell^* + \mathbf{d}_\ell^{t+1}$. Now, define the following distance function:

$$d_\infty([U, \Sigma], [U^*, \Sigma^*]) \ \equiv \ \max_\ell \left(\|\mathbf{d}_\ell\|_2 + \Delta_\ell^\sigma \right) \ .$$

The next theorem shows that this distance function decreases geometrically with number of iterations of Algorithm 1. A proof of this theorem is provided in Appendix B.4.

Theorem 2.3. *If $d_\infty([U, \Sigma], [U^*, \Sigma^*]) \leq \frac{1}{1600r} \frac{\sigma_{min}^*}{\sigma_{max}^*}$ and $\mathbf{u}_i$ is 2μ-incoherent for all $1 \leq i \leq r$, then there exists a positive constant C such that for $p \geq \frac{Cr^2(\sigma_{\max}^*)^2 \mu^3 \log^2 n}{(\sigma_{\min}^*)^2 n^{3/2}}$ we have w.p. $\geq 1 - \frac{1}{n^7}$,*

$$d_\infty([U^{t+1}, \Sigma^{t+1}], [U^*, \Sigma^*]) \leq \frac{1}{2} d_\infty([U, \Sigma], [U^*, \Sigma^*]),$$

where $[U^{t+1}, \ \Sigma^{t+1}] = \{(\mathbf{u}_\ell^{t+1}, \sigma_\ell^{t+1}), 1 \leq \ell \leq r\}$ are the $(t+1)$-th step iterates of Algorithm 1. Moreover, each $\mathbf{u}_\ell^{t+1}$ is 2μ-incoherent for all ℓ.
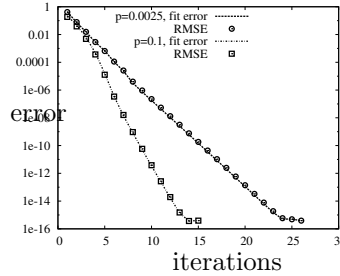

Figure 2: Algorithm 1 exhibits linear convergence until machine precision. For the estimate $\widehat{T}_t$ at the t-th iterations, the fit error $\|\mathcal{P}_\Omega(T - \widehat{T}_t)\|_F / \|\mathcal{P}_\Omega(T)\|_F$ closely tracks the normalized root mean squared error $\|T - \widehat{T}_t\|_F / \|T\|_F$, suggesting that it serves as a good stopping criterion.

Note that our number of samples depend on the number of iterations τ. But due to linear convergence, our sample complexity increases only by a factor of $\log(1/\epsilon)$ where ϵ is the desired accuracy.

Difference from Matrix AltMin: Here, we would like to highlight differences between our analysis and analysis of the alternating minimization method for matrix completion (matrix AltMin) [4, 5]. In the matrix case, the singular vectors $\mathbf{u}_i^*$'s need not be unique. Hence, the analysis is required to guarantee a decay in the subspace distance $dist(U, U^*)$; typically, principal angle based subspace distance is used for analysis. In contrast, orthonormal $\mathbf{u}_i^*$'s uniquely define the tensor and hence one can obtain distance bounds $\|\mathbf{u}_i - \mathbf{u}_i^*\|_2$ for each component $\mathbf{u}_i$ individually.

On the other other hand, an iteration of the matrix AltMin iterates over all the vectors $\mathbf{u}_i, 1 \leq i \leq r$, where r is the rank of the current iterate and hence don't have to consider the error in estimation of the fixed components $U_{[r]\backslash q} = \{\mathbf{u}_\ell, \ \forall \ \ell \neq q\}$, which is a challenge for the analysis of Algorithm 1 and requires careful decomposition and bounds of the error terms.

7

2.3 Proof of Theorem 1.1

Let $T = \sum_{q=1}^{r} \sigma_q^* (u_q^* \otimes u_q^* \otimes u_q^*)$. Denote the initial estimates $U^0 = [u_1^0, \ldots, u_r^0]$ and $\sigma^0 = [\sigma_1^0, \ldots, \sigma_r^0]$ to be the output of robust tensor power method at step 5 of Algorithm 1. With a choice of $p \geq C(\sigma_{\max}^*)^4 \mu^6 r^4 (\log n)^4 / (\sigma_{\min}^*)^4 n^{3/2}$ as per our assumption, Lemma 2.2 ensures that we have $\|u_q^0 - u_q^*\| \leq \sigma_{\min}^* / (4800\, r\sigma_{\max})$ and $|\sigma_q^0 - \sigma_q^*| \leq |\sigma_q^*| \sigma_{\min}^* / (4800\, r\sigma_{\max})$ with probability at least $1 - n^{-5}$. This requires running robust tensor power method for $(r \log n)^c$ random initializations for some positive constant c, each requiring $O(|\Omega|)$ operations ignoring logarithmic factors.

To ensure that we have sufficiently incoherent initial iterate, we perform thresholding proposed in [4]. In particular, we threshold all the elements of $\mathbf{u}_i^0$ (obtained from RTPM method, see Step 3 of Algorithm 1) that are larger (in magnitude) than $\mu/\sqrt{n}$ to be $\frac{sign(\mathbf{u}_\ell(i))\mu}{\sqrt{n}}$ and then re-normalize to obtain $\mathbf{u}_i$. Using Lemma A.4, this procedure ensures that the obtained initial estimate $\mathbf{u}_i$ satisfies the two criteria that is required by Theorem 2.3: a) $\|\mathbf{u}_i - \mathbf{u}_i^*\|_2 \leq \frac{1}{1600r} \cdot \frac{\sigma_{min}^*}{\sigma_{max}^*}$, and b) $\mathbf{u}_i$ is 2μ-incoherent.

With this initialization, Theorem 2.3 tells us that $O(\log_2(4r^{1/2}\|T\|_F/\varepsilon))$ iterations (each iteration requires $O(r|\Omega|)$ operations) is sufficient to achieve:

$$\|u_q - u_q^*\|_2 \leq \frac{\varepsilon}{4r^{1/2}\|T\|_F} \quad \text{and} \quad |\sigma_q - \sigma_q^*| \leq \frac{|\sigma_q^*|\varepsilon}{4r^{1/2}\|T\|_F} \ ,$$

for all $q \in [r]$ with probability at least $1 - n^{-7} \log_2(4r^{1/2}\|T\|_F/\varepsilon)$. The desired bound follows from the next lemma with a choice of $\tilde{\varepsilon} = \varepsilon/4r^{1/2}\|T\|_F$. For a proof we refer to Appendix B.6.

Lemma 2.4. *For an orthogonal rank-r tensor $T = \sum_{q=1}^{r} \sigma_q^*(u_q^* \otimes u_q^* \otimes u_q^*)$ and any rank-r tensor $\widehat{T} = \sum_{q=1}^{r} \sigma_q(u_q \otimes u_q \otimes u_q)$ satisfying $\|u - u^*\|_2 \leq \tilde{\varepsilon}$ and $|\sigma - \sigma^*| \leq |\sigma^*|\tilde{\varepsilon}$ for all $q \in [r]$ and for all positive $\tilde{\varepsilon} > 0$, we have $\|T - \widehat{T}\|_F \leq 4\, r^{1/2}\, \|T\|_F\, \tilde{\varepsilon}$.*

2.4 Fundamental limit and random hypergraphs

For matrices, it is known that exact matrix completion is impossible if the underlying graph is disconnected. A refinement of this analysis for Erdös-Renýi graphs provides a lower bound on matrix completion: when sample size is less than $C\mu rn \log n$, no algorithm can recover the original matrix [6].

However, for tensor completion and random hyper graphs, such a simple connection does not exist. It is not known how the properties of the hyper graph is related to recovery. In this spirit, a rank-one third-order tensor completion has been studied in a specific context of *MAX-3LIN problems*. Consider a series of linear equations over n binary variables $x = [x_1 \ldots x_n] \in \{\pm 1\}^n$. An instance of a 3LIN problem consists of a set of linear equations on GF(2), where each equation involve exactly three variables, e.g.

$$
\begin{aligned}
x_1 \oplus x_2 \oplus x_3 &= +1 \\
x_2 \oplus x_3 \oplus x_4 &= -1 \\
x_3 \oplus x_4 \oplus x_5 &= +1
\end{aligned}
\tag{6}
$$

We use -1 to denote true (or 1 in GF(2)) and $+1$ to denote false (or 0 in GF(2)). Then the exclusive-or operation denoted by $\oplus$ is the integer multiplication. the MAX-3LIN problem is to find a solution x that satisfies as many number of equations as possible. This is an NP-hard problem in general, and hence random instances of the problem with a *planted solution* has been studied [23]. Algorithm 1 provides a provable guarantee for MAX-3LIN with random assignments.

Corollary 2.5. *For random MAX-3LIN problem with a planted solution, under the hypotheses of Theorem 1.1, Algorithm 1 finds the correct solution with high probability.*

Notice that the tensor has incoherence one and rank one. This implies exact reconstruction for $P \geq C(\log n)^4/n^{3/2}$. This significantly improves over a message-passing approach to MAX-3LIN in [23], which is guaranteed to find the planted solution for $p \geq C(\log \log n)^2/(n \log n)$. It was suggested that a new notion of connectivity called *propagation connectivity* is a sufficient condition for the solution of random MAX-3LIN problem with a planted solution to be unique [23, Proposition 2]. Precisely, it is claimed that if the hypergraph corresponding to an instance of MAX-3LIN is propagation connected, then the optimal solution for MAX-3LIN is unique and there is an efficient algorithm that finds it. However, the example in 6 is propagation connected but there is no unique solution: $[1, 1, 1, -1, -1], [1, -1, -1, 1, -1], [-1, 1, -1, 1, -1], [-1, -1, 1, 1, 1]$ all satisfy the equations and corresponding tensor $T = x \otimes x \otimes x$ is not uniquely defined either. This proves that propagation connectivity is not a sufficient condition for uniqueness of the MAX-3LIN solution.

2.5 Open Problems and Future Directions

Tensor completion for non-orthogonal decomposition. Numerical simulations suggests that non-orthogonal CP models can be recovered exactly (without the usual whitening step). It would be interesting to analyze our algorithm under non-orthogonal CP model. However, we would like to point here that even with fully observed tensor, exact factorization is known only for orthonormal tensors. Now, given that our method guarantees not only completion but also factorization of the tensor (which is essential for large scale applications), it is natural that our method would require a similar condition.

Optimal dependence on r. The numerical results suggests threshold sample size scaling as $\sqrt{r}$. This is surprising since the degrees of freedom in describing CP model scales as linearly in r. This implies that the $\sqrt{r}$ scaling can only hold for $r = O(\sqrt{n})$. In comparison, for matrix completion we know that the threshold scales as r. It would be important to understand why this change in dependence in r happens for higher order tensors, and identify how it depends on k for k-th order tensor completion.

References

[1] Anandkumar Anima, Ge Rong, Hsu Daniel, M. Kakade Sham, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.

[2] A. Anandkumar, R. Ge, and M. Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[3] V. De Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[4] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674, 2013.

[5] Moritz Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint arXiv:1312.0925*, 2013.

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

[7] F. L. Hitchcock. *The expression of a tensor or a polyadic as a sum of products*. sn., 1927.

[8] J. D. Carroll and J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319, 1970.

[9] R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[10] T. Zhang and G. H. Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23(2):534–550, 2001.

[11] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.

[12] G. Tomasi and R. Bro. Parafac and missing values. *Chemometrics and Intelligent Laboratory Systems*, 75(2):163–180, 2005.

[13] R. Bro. *Multi-way analysis in the food industry: models, algorithms, and applications*. PhD thesis, 1998.

[14] B Walczak and DL Massart. Dealing with missing data: Part i. *Chemometrics and Intelligent Laboratory Systems*, 58(1):15–27, 2001.

[15] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[16] C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*, 2013.

[17] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[18] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980–2998, June 2010.

[19] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *NIPS*, pages 1331–1339, 2013.

[20] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 619–626. ACM, 2001.

[21] J. Friedman, J. Kahn, and E. Szemerédi. On the second eigenvalue in random regular graphs. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 587–598, Seattle, Washington, USA, may 1989. ACM.

[22] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Struct. Algorithms*, 27(2):251–275, 2005.

[23] R. Berke and M. Onsjö. Propagation connectivity of random hypergraphs. In *Stochastic Algorithms: Foundations and Applications*, pages 117–126. Springer, 2009.

A Proof of Theorem 2.1 for Initialization Analysis

We prove the following bound on the spectrum of random tensors:

$$\max_{x,y,z,\|x\|=\|y\|=\|z\|=1} \left(\mathcal{P}_\Omega(T) - p\,T \right)[x,y,z] \quad \leq \quad C\,T_{\max}\,(\log n)^2\,\sqrt{(n_1\,n_2\,n_3)^{1/2}\,p}\,.$$

Here we prove the theorem for general case where T is not symmetric and might even have different dimensions n_1, n_2 and n_3. Inspired by [21, 18], our strategy is as follows:

(1) Reduce to x,y, and z which belongs to *discretized sets* $\widetilde{S}_{n_1}$, $\widetilde{S}_{n_2}$, and $\widetilde{S}_{n_3}$;

(2) Bound the contribution of *light triples* using concentration of measure;

(3) Bound the contribution of *heavy triples* using the discrepancy property of a random tripartite hypergraph.

Define a discretization of an n-dimensional ball as

$$\widetilde{S}_n \ \equiv \ \left\{ x \in \left\{ \frac{\Delta}{\sqrt{n}} \mathbb{Z} \right\}^n \ : \ \|x\| \leq 1 \right\},$$

such that $\widetilde{S}_n \subseteq S_n \equiv \{ x \in \mathbb{R}^n \ : \ \|x\| \leq 1 \}$. Later we will set Δ to be a small enough constant.

Lemma A.1 (Remark 4.1 in [18])**.** *For any tensor $A \in \mathbb{R}^{n_1 \times n_2 \times n_3}$,*

$$\max_{x \in S_{n_1}, y \in S_{n_2}, x \in S_{n_3}} A[x, y, z] \ \leq \ \max_{x \in \widetilde{S}_{n_1}, y \in \widetilde{S}_{n_2}, x \in \widetilde{S}_{n_3}} \frac{1}{(1 - \Delta)^3} A[x, y, z]$$

It is therefore enough show that the bound holds for discretized vectors all discretized vectors x, y, and z. One caveat is that such a probabilistic bound must hold with probability sufficiently close to one such that we can apply the union bound over all discretized choices of x, y, and z. The following lemma bounds the number of such choices.

Lemma A.2 ([18])**.** *The size of the discretized set is bounded by $|\widetilde{S}_n| \leq (\Delta/10)^n$.*

A naive approach to upper bound $(\mathcal{P}_\Omega(T) - p\,T)[x, y, z]$ would be to consider it as a random variable and apply concentration inequalities directly. However, this naive approach fails since x, y and z can contain entries that are much larger than their typical value of $O(1/\sqrt{n})$. We thus separate the analysis into two contributions, and apply concentration inequalities to bound the contribution of the *light triples* and use graph topology of the random sampling to bound the contribution of the *heavy triples*. Define the light triples as

$$\mathcal{L} \ \equiv \ \left\{ (i, j, k) \ : \ \left| T_{ijk} x_i y_j z_k \right| \ \leq \ T_{\max} \sqrt{\frac{\epsilon}{n_1 n_2 n_3}} \right\}. \tag{7}$$

Heavy triples are defined as its complement $\overline{\mathcal{L}} = \{[n_1] \times [n_2] \times [n_3]\} \setminus \mathcal{L}$. Later we will set the appropriate value for $\epsilon = \Theta(p\sqrt{n_1 n_2 n_3})$. We can then write each contributions separately as

$$\left| (\mathcal{P}_\Omega(T) - p\,T)[x, y, z] \right| \ \leq \ \left| \sum_{(i,j,k) \in \mathcal{L}} \left(\mathcal{P}_\Omega(T)_{ijk} x_i y_j z_k \right) - p\,T[x, y, z] \right| + \left| \sum_{(i,j,k) \in \overline{\mathcal{L}}} \mathcal{P}_\Omega(T)_{ijk} x_i y_j z_k \right|. \tag{8}$$

We will prove that both contributions are upper bounded by $CT_{\max}(\log n)^2 \sqrt{(n_1 n_2 n_3)^{1/2} p}$ with some positive constant C for all $x \in \widetilde{S}_{n_1}$, $y \in \widetilde{S}_{n_2}$, and $z \in \widetilde{S}_{n_3}$. The bound on the light triples follows from Chernoff's concentration inequalities. The bound on the heavy triples follows from the discrepancy property of random hyper graphs, which implies that there cannot be too many triples with large contributions. Theorem 2.1 then follows from Lemma A.1 with an appropriate choice of $\Delta = \Theta(1)$.

A.1   Bounding the contribution of light triples

Let $Z \equiv \sum_{(i,j,k) \in \mathcal{L}} \left(\mathcal{P}_\Omega(T)_{ijk} x_i y_j z_k \right) - p\,T[x, y, z]$ for some $x \in S_{n_1}$, $y \in S_{n_2}$, and $z \in S_{n_3}$. We claim that

$$\mathbb{P}\left(Z > \frac{p T_{\max} \sqrt{n_1 n_2 n_3}}{\sqrt{\epsilon}} + t\,(n_1 + n_2 + n_3) \frac{2 T_{\max} \sqrt{\epsilon}}{\sqrt{n_1 n_2 n_3}} \right) \ \leq \ \exp\left\{ - t(n_2 + n_2 + n_3) \right\}. \tag{9}$$

We first show that the mean of Z is bounded as

$$\left| \mathbb{E}[Z] \right| \ \leq \ 2\, p\, T_{\max} \sqrt{\frac{n_1 n_2 n_3}{\epsilon}}. \tag{10}$$

The mean can be written as $\mathbb{E}[Z] = p\sum_{\mathcal{L}} T_{ijk}x_iy_jz_k - p\sum_{[n_1]\times[n_2]\times[n_3]} T_{ijk}x_iy_jz_k = p\sum_{\overline{\mathcal{L}}} T_{ijk}x_iy_jz_k$. Using the fact that for heavy triples $|T_{ink}x_iy_jz_k| \geq T_{\max}\sqrt{\epsilon/(n_1n_2n_3)}$, the expected contribution is then bounded by

$$
\begin{aligned}
\Big| \sum_{(i,j,k)\in\overline{\mathcal{L}}} T_{ijk}x_iy_jz_k \Big| &\leq \sum_{(i,j,k)\in\overline{\mathcal{L}}} \frac{T_{ijk}^2 x_i^2 y_j^2 z_k^2}{|T_{ijk}x_iy_jz_k|} \\
&\leq \frac{\sqrt{n_1n_2n_3}}{T_{\max}\sqrt{\epsilon}} \sum_{(i,j,k)\in\overline{\mathcal{L}}} T_{ijk}^2 x_i^2 y_j^2 z_k^2 \\
&\leq \frac{T_{\max}\sqrt{n_1n_2n_3}}{\sqrt{\epsilon}}\ .
\end{aligned}
$$

We next show concentration of Z around item mean. Let $\lambda = \sqrt{n_1n_2n_3}/(2T_{\max}\sqrt{\epsilon})$ such that $|\lambda T_{ijk}x_iy_jz_k| \leq 1/2$ for all $(i,j,k)\in\mathcal{L}$. Then, $e^{\lambda T_{ijk}x_iy_jz_k} - 1 \leq \lambda T_{ijk}x_iy_jz_k + 2\lambda(T_{ijk}x_iy_jz_k)^2$.

$$
\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &= \exp\{-\lambda\,p\,T[x,y,z]\} \prod_{(i,j,k)\in\mathcal{L}} \big(1-p+p\,e^{\lambda T_{ijk}x_iy_jz_k}\big) \\
&\leq \exp\{-\lambda\,p\,T[x,y,z]\} \prod_{(i,j,k)\in\mathcal{L}} \big(1+p\,(\lambda T_{ijk}x_iy_jz_k + 2\lambda^2(T_{ijk}x_iy_jz_k)^2)\big) \\
&\leq \exp\Big\{p\sum_{\mathcal{L}}\lambda T_{ijk}x_iy_jz_k - \lambda\,p\,T[x,y,z] + p\sum_{\mathcal{L}} 2\lambda^2(T_{ijk}x_iy_jz_k)^2\Big\} \\
&\leq \exp\Big\{\lambda\mathbb{E}[Z] + \frac{p\,n_1n_2n_3}{2\epsilon}\Big\}\ .
\end{aligned}
$$

Applying Chernoff bound $\mathbb{P}(Z - \mathbb{E}[Z] > t) \leq \mathbb{E}[e^{\lambda Z}]e^{-\lambda\mathbb{E}[Z]-\lambda t}$, this proves (9). Note that the deviation of $-Z$ can be bounded similarly. We can now finish the proof of upper bound on the contribution of light triples by taking the union bound over all discretized vectors inside the ball. Setting $t = 2\log(20/\Delta)$ in (9), we get

$$
\mathbb{P}\Big(\max_{x\in\widetilde{S}_{n_1},y\in\widetilde{S}_{n_2},z\in\widetilde{S}_{n_3}} \sum_{(i,j,k)\in\mathcal{L}} T_{ijk}x_iy_jz_k - p\,T[x,y,z] \geq \frac{pT_{\max}\sqrt{n_1n_2n_3}}{\sqrt{\epsilon}} + 2\log(20/\Delta)\,(n_1+n_2+n_3)\,\frac{2T_{\max}\sqrt{\epsilon}}{\sqrt{n_1n_2n_3}}\Big)
$$

$$
\leq 2\,e^{(n_1+n_2+n_3)\log(20/\Delta)}\,e^{-2\log(20/\Delta)(n_1+n_2+n_3)}
$$

$$
\leq 2e^{-(n_1+n_2+n_3)\log(20/\Delta)}\ .
$$

Since $p = \epsilon/\sqrt{n_1n_2n_3}$, this proves that the contribution of light triples is bounded by $C\,T_{\max}\sqrt{\epsilon}$ with high probability.

Note that for the range of $p = \epsilon/n^2$, the contribution of light couples is bounded by $C\,T_{\max}\sqrt{\epsilon/n}$. However, even in this regime of p, the contribution of heavy triples is still $\Omega(1)$, which dominates the light triples by a factor of $\sqrt{n}$. This is the reason for the choice of $p = \Theta(\epsilon/n^{1.5})$.

A.2 Bounding the contribution of heavy triples

The contribution of heavy triples is bounded by

$$
\Big| \sum_{(i,j,k)\in\overline{\mathcal{L}}} T_{ijk}x_iy_jz_k \Big| \leq T_{\max} \sum_{(i,j,k)\in\overline{\mathcal{L}}} |x_iy_jz_k|\ .
$$

In the following, we will show that the right-hand side of the above inequality is upper bounded by

$$
\sum_{(i,j,k)\in\overline{\mathcal{L}}} |x_iy_jz_k| \leq C\sqrt{\epsilon}(\log n)^2\ ,
$$

12

for some positive numerical constant $C > 0$ with probability larger than $1 - n^{-5}$.

We consider a hypergraph $G = ([n_1] \times [n_2] \times [n_3], E)$ with undirected hyper edges, where each edge connects three nodes, each one from each set $[n_1]$, $[n_2]$, and $[n_3]$. Given a sampling of entries in a tensor, we let the edges in G denote the positions of the entries that is sampled. The proof is a generalization of similar proof for matrices in [21, 22, 18] and is based on two properties of the hypergraph G. Define the degree of a node as the number of edges connected to that particular node such that $\deg_1(i) \equiv |\{(i, j, k) \in E\}|$, and similarly define $\deg_2(j)$ and $\deg_3(k)$. Define the degree of two nodes as the number of edges connected to both of the nodes such that $\deg_{12}(i, j) \equiv |\{(i, j, k) \in E\}|$, and similarly define $\deg_{13}(i, k)$ and $\deg_{23}(j, k)$.

1. *Bounded degree property.* A hyper graph G satisfies the bounded degree property if the degree are upper bounded as follows:

$$
\begin{aligned}
\deg_1(i) &\leq \xi_0 \, p \, n_2 n_3 && \text{for all } i \in [n_1] \,, \\
\deg_2(j) &\leq \xi_0 \, p \, n_1 n_3 && \text{for all } j \in [n_2] \,, \\
\deg_3(k) &\leq \xi_0 \, p \, n_1 n_2 && \text{for all } k \in [n_3] \,, \\
\deg_{12}(i, j) &\leq \xi_0 \, (p \, n_3 + \log n_3) && \text{for all } i \in [n_1], j \in [n_2] \,, \\
\deg_{13}(i, k) &\leq \xi_0 \, (p \, n_2 + \log n_2) && \text{for all } i \in [n_1], k \in [n_3] \,, \\
\deg_{23}(j, k) &\leq \xi_0 \, (p \, n_1 + \log n_1) && \text{for all } j \in [n_2], k \in [n_3] \,,
\end{aligned}
\tag{11}
$$

for some positive numerical constant $\xi_0 > 0$ (independent of n_1, n_2, n_3 and p) where $p = |E|/(n_1 n_2 n_3)$.

2. *Discrepancy property.* A hyper graph G satisfies the discrepancy property if for any subset of nodes $A_1 \in [n_1]$, $A_2 \in [n_2]$, and $A_3 \in [n_3]$, at least one of the following is true:

$$
e(A_1, A_2, A_3) \leq \xi_1 \, \bar{e}(A_1, A_2, A_3) \,,
\tag{12}
$$

$$
e(A_1, A_2, A_3) \ln \left(\frac{e(A_1, A_2, A_3)}{\bar{e}(A_1, A_2, A_3)} \right) \leq \xi_2 \, \max \left\{ |A_1| \ln \left(\frac{e \, n_1}{|A_1|} \right) , \, |A_2| \ln \left(\frac{e \, n_2}{|A_2|} \right) , \, |A_3| \ln \left(\frac{e \, n_3}{|A_3|} \right) \right\}
\tag{13}
$$

for some positive numerical constants $\xi_1, \xi_2 > 0$ (independent of n_1, n_2, n_3 and p). Here, $e(A_1, A_2, A_3)$ denotes the number of edges between the three subsets A_1, A_2 and A_3, and $\bar{e}(A_1, A_2, A_3) \equiv p \, |A_1| \, |A_2| \, |A_3|$ denotes the average number of edges between the three subsets.

We first prove that if the sampling pattern is defined by a graph G which satisfies both the bounded degree and discrepancy properties, then the contribution of heavy triples is $O(\sqrt{\epsilon})$. Notice that this is a deterministic statement, that holds for all graphs with the above properties. We then finish the proof by showing that the random sampling satisfies both the bounded degree and discrepancy properties with probability at least $1 - n^{-5}$.

We partition the indices according to the value of corresponding vectors:

$$
\begin{aligned}
A_1^{(u)} &\equiv \left\{ i \in [n_1] : \frac{\Delta}{\sqrt{n_1}} 2^{u-1} \leq |x_i| < \frac{\Delta}{\sqrt{n_1}} 2^u \right\} \\
A_2^{(v)} &\equiv \left\{ j \in [n_2] : \frac{\Delta}{\sqrt{n_2}} 2^{v-1} \leq |y_j| < \frac{\Delta}{\sqrt{n_2}} 2^v \right\} \\
A_3^{(w)} &\equiv \left\{ k \in [n_3] : \frac{\Delta}{\sqrt{n_3}} 2^{w-1} \leq |z_k| < \frac{\Delta}{\sqrt{n_3}} 2^w \right\} \,,
\end{aligned}
$$

for $u \in \{1, \ldots, \lceil \log_2(\sqrt{n_1}/\Delta) \rceil + 1\}$, $v \in \{1, \ldots, \lceil \log_2(\sqrt{n_2}/\Delta) \rceil + 1\}$, and $w \in \{1, \ldots, \lceil \log_2(\sqrt{n_3}/\Delta) \rceil + 1\}$. We denote the size of each set by $a_i^{(u)} \equiv |A_i^{(u)}|$. We use e_{uvw} to denote the number of edges between three subsets $A_1^{(u)}$, $A_2^{(v)}$, and $A_3^{(w)}$, and we use $\bar{e}_{uvw} \equiv p \, a_1^{(u)} a_2^{(v)} a_3^{(w)}$ to denote the average number of edges. Notice that the above definition of $A_1^{(u)}$'s cover all non-zero values of the entries of x, since, with discretization, the smallest possible positive value is $\Delta/\sqrt{n_1}$. The same applies to the entries of y and z.

13

$$\sum_{(i,j,k)\in\overline{\mathcal{L}}} \left|x_i y_j z_k\right| \;\leq\; \sum_{(i,j,k):|x_i y_j z_k|>\sqrt{\epsilon/(n_1 n_2 n_3)}} \left|x_i y_j z_k\right|$$

$$\leq \sum_{(u,v,w):2^{u+v+w}>8\sqrt{\epsilon}/\Delta^3} e_{uvw} \underbrace{\frac{\Delta 2^u}{\sqrt{n_1}}\frac{\Delta 2^v}{\sqrt{n_2}}\frac{\Delta 2^w}{\sqrt{n_3}}}_{\sigma_{uvw}}\;.$$

Note that since $\sum_u a_1^{(u)} 2^{2(u-1)}\Delta^2/n_1 \leq \|x\|^2 \leq 1$, we get that

$$
\begin{aligned}
a_1^{(u)} &\leq (n_1/\Delta^2)2^{-2(u-1)}\;,\\
a_2^{(v)} &\leq (n_2/\Delta^2)2^{-2(v-1)}\;,\\
a_3^{(w)} &\leq (n_3/\Delta^2)2^{-2(w-1)}\;.
\end{aligned}
\tag{14}
$$

The contributions from various combinations of (u,v,w) utilize various subsets of our assumptions. We prove that in each case the contribution is $O(\sqrt{\epsilon}(\log n)^2)$ as follows.

Case1. For (u,v,w) satisfying the first discrepancy property (12) : $e_{uvw} \leq \xi_1 \bar{e}_{uvw}$.

In this case, using (14) and the fact that $p = \epsilon/\sqrt{n_1 n_2 n_3}$,

$$
\begin{aligned}
\sum \sigma_{uvw} &\leq \xi_1 p a_1^{(u)} a_2^{(v)} a_3^{(w)} \frac{\Delta^3 2^{u+v+w}}{\sqrt{n_1 n_2 n_3}}\\
&\leq \frac{64\xi_1 p \sqrt{n_1 n_2 n_3}}{\Delta^3 2^{u+v+w}}\\
&\leq 16\xi_1\sqrt{\epsilon}(\log n)^2\;,
\end{aligned}
$$

where $n \equiv \max\{n_1,n_2,n_3\}$ and in the last inequality we used the fact that we are summing over heavy triples satisfying $\Delta^3 2^{u+v+w} > 8\sqrt{\epsilon}$, and $\sum_{(u,v,w):2^{u+v+w}\leq 8\sqrt{\epsilon}/\Delta^3} 2^{-(u+v+w)} \leq 2\log_2(\sqrt{n_1}/\Delta)\log_2(\sqrt{n_2}/\Delta)\Delta^3/(8\sqrt{\epsilon})$.

Case2. For (u,v,w) satisfying the second discrepancy property in (13).

Case 2-1. For (u,v,w) satisfying $\ln(e_{uvw}/\bar{e}_{uvw}) \leq (1/2)\ln(en_3/a_3^{(w)}) = (1/4)(\ln(en_3/(a_3^{(w)}2^{2w}))+\ln(2^{2w}))$.

Case 2-1-1. When $\ln(2^{2w}) \leq \ln(en_3/(a_3^{(w)}2^{2w}))$, we have $\ln(e_{uvw}/\bar{e}_{uvw}) \leq \ln(en_2/(a_3^{(w)}2^{2w}))$, which gives

$$
\begin{aligned}
e_{uvw} &\leq en_3 \bar{e}_{uvw}/(a_3^{(w)}2^{2w})\\
&\leq e\, a_1^{(u)} a_2^{(v)} n_3 p 2^{-2w}\\
&\leq \frac{16e}{\Delta^4}\, n_1 n_2 n_3 p 2^{-2(u+v+w)}\;.
\end{aligned}
$$

It follows that $\sum \sigma_{uvw} \leq (16/\Delta)p\sqrt{n_1 n_2 n_3}2^{-u-v-w} \leq 2\Delta^2\sqrt{\epsilon}(\log n)^2$ using the fact that we are summing over heavy triples.

Case 2-1-2. When $\ln(2^{2w}) > \ln(en_3/(a_3^{(w)}2^{2w}))$, we have $\ln(e_{uvw}/\bar{e}_{uvw}) \leq \ln(2^w)$.

Case 2-1-2-1. For $\sqrt{\epsilon}e_{uvw} > 2^{u+v+w}\bar{e}_{uvw}$, it follows that $2^{u+v} \leq \sqrt{\epsilon}$. Since we are in the case where the first discrepancy does not hold, i.e. $e_{uvw} > \xi_1 \bar{e}_{uvw}$, and the the second discrepancy property holds, we have $e_{uvw} \leq e_{uvw}\ln(e_{uvw}/\bar{e}_{uvw}) \leq \xi_2 a_3^{(w)}\ln(en_3/a_3^{(w)}) \leq$

$2\xi_2 a_3^{(w)} \ln(2^{2w})$ Then,

$$\sum \sigma_{uvw} \leq \sum 2\xi_2 a_3^{(w)} \ln(2^{2w}) \frac{\Delta^3 2^{u+v+w}}{\sqrt{n_1 n_2 n_3}}$$

$$\leq \sum \frac{8\xi_2 \Delta \sqrt{n_3} 2^{u+v}}{\sqrt{n_1 n_2}} \frac{\ln(2^w)}{2^w}$$

$$\leq 8\Delta\xi_2 \sqrt{\frac{n_3}{n_1 n_2}} \sqrt{\epsilon} \log_2(\sqrt{n_1}/\Delta) \log_2(\sqrt{n_3}/\Delta) \, ,$$

which is $O\big(\sqrt{\epsilon}(\log n)^2 \sqrt{(n_3/(n_1 n_2))}\,\big)$

Case 2-1-2-2. For $\sqrt{\epsilon} e_{uvw} \leq 2^{u+v+w} \bar{e}_{uvw}$,

$$\sum \sigma_{uvw} \leq \sum \frac{2^{u+v+w} p\, a_1^{(u)} a_2^{(v)} a_3^{(w)}}{\sqrt{\epsilon}} \frac{\Delta^3 2^{u+v+w}}{\sqrt{n_1 n_2 n_3}}$$

$$\leq \frac{\sqrt{\epsilon}}{\Delta^3} \sum \frac{a_1^{(u)} a_2^{(v)} a_3^{(w)} \Delta^3 2^{2(u+v+w)}}{n_1 n_2 n_3}$$

$$\leq \frac{\sqrt{\epsilon}}{\Delta^3} \|x\|^2 \|y\|^2 \|z\|^2 \, ,$$

which is $O(\sqrt{\epsilon})$.

Case 2-2. For (u,v,w) satisfying $\ln(e_{uvw}/\bar{e}_{uvw}) > (1/2)\ln(en_3/a_3^{(w)})$.

Case 2-2-1. For $2^{u+v} \leq \sqrt{n_1 n_2 \epsilon / n_3} 2^w$, we know from the condition $\ln(e_{uvw}/\bar{e}_{uvw}) > (1/2)\ln(en_3/a_3^{(w)})$, that $e_{uvw} \leq 2\xi_2 a_3^{(w)}$. Then,

$$\sum \sigma_{uvw} \leq \sum 2\xi_2 a_3^{(w)} \frac{2^{u+v+w}\Delta^3}{\sqrt{n_1 n_2 n_3}}$$

$$\leq \sum 8\xi_2 \Delta 2^{u+v-w} \sqrt{\frac{n_3}{n_1 n_2}}$$

$$\leq 8\xi_2 \Delta \sqrt{\epsilon} \log_2(\sqrt{n_1}/\Delta) \log_2(\sqrt{n_2}/\Delta) \, ,$$

which is $O(\sqrt{\epsilon}(\log n)^2)$.

Case 2-2-2. For $2^{u+v} > \sqrt{n_1 n_2 \epsilon / n_3} 2^w$

Case 2-2-2-1. For (u,v,w) satisfying bounded degree property with $\deg_{12}(i,j) \leq \xi_0 p n_3$, we have $e_{uvw} \leq a_1 a_2 \xi_0 p n_3$. Then,

$$\sum \sigma_{uvw} \leq \sum \frac{16\xi_0 \epsilon}{\Delta} 2^{w-u-v}$$

$$\leq \sqrt{\frac{n_3}{n_1 n_2}} \frac{16\xi_0 \sqrt{\epsilon}}{\Delta} \log_2(\sqrt{n_2}/\Delta) \log_2(\sqrt{n_2}/\Delta) \, ,$$

which is $O(\sqrt{\epsilon}\sqrt{n_3/(n_1 n_2)}(\log_2 n)^2)$.

Case 2-2-2-2. For (u,v,w) satisfying bounded degree property with $\deg_{12}(i,j) \leq \xi_0 \log n_3$, we have $e_{uvw} \leq a_1 a_2 \xi_0 \log n$.

$$\sum \sigma_{uvw} \leq \sum \frac{16\xi_0 \log n_3 2^{w-u-v}}{\Delta} \sqrt{\frac{n_1 n_2}{n_3}}$$

$$\leq \frac{16\xi_0 \log n_3}{\Delta\sqrt{\epsilon}} \log_2(\sqrt{n_1}/\Delta) \log_2(\sqrt{n_2}/\Delta) \, ,$$

which is $O((1/\sqrt{\epsilon})(\log n)^3)$.

15

For $\epsilon \geq \log n$, this proves that the contribution of the heavy triples is $O(\sqrt{\epsilon}(\log n)^2)$.

We are left to prove that the bounded degree and the bounded discrepancy properties hold for a random tripartite hypergraph $G = (V_1 \cup V_2 \cup V_3, E)$ where each edge is selected with probability p. Precisely, let $n = \max\{|V_1|, |V_2|, |V_3|\}$, then the following lemma provides a bound on the degree and discrepancy, with high probability.

Lemma A.3. *For any $\delta \in [0, 1/e]$ and $p \geq (1/n^2) \log n$, there exists numerical constants $C, C' > 0$ such that a random tripartite hyper graph satisfies the bounded degree property: for all $i \in V_1$, $j \in V_2$, and $k \in V_3$,*

$$\deg_1(i) \leq 2pn_2n_3 + \frac{8}{3}\log\frac{3n_1}{\delta}$$

$$\deg_2(j) \leq 2pn_1n_3 + \frac{8}{3}\log\frac{3n_2}{\delta}$$

$$\deg_3(k) \leq 2pn_1n_2 + \frac{8}{3}\log\frac{3n_3}{\delta}$$

$$\deg_{12}(i,j) \leq 2pn_3 + \frac{8}{3}\log\frac{3n_1n_2}{\delta}$$

$$\deg_{13}(i,k) \leq 2pn_2 + \frac{8}{3}\log\frac{3n_1n_3}{\delta}$$

$$\deg_{23}(j,k) \leq 2pn_1 + \frac{8}{3}\log\frac{3n_2n_3}{\delta}$$

and the bounded discrepancy property: for all subsets $A_1 \subseteq V_1$, $A_2 \subseteq V_2$, and $A_3 \subseteq V_3$, at least one of the following is true.

$$e(A_1, A_2, A_3) \leq C\alpha^2 \bar{e}(A_1, A_2, A_3)\left(1 + \frac{\ln(1/\delta)}{pn^2}\right), \quad or$$

$$e(A_1, A_2, A_3)\ln\left(\frac{e(A_1, A_2, A_3)}{\bar{e}(A_1, A_2, A_3)}\right) \leq C'\left(\ln\frac{\alpha}{\delta} + \max\left\{|A_1|\ln\frac{e\,n_1}{|A_1|}, \, |A_2|\ln\frac{e\,n_2}{|A_2|}, \, |A_3|\ln\frac{e\,n_3}{|A_3|}\right\}\right),$$

where $n_1 = |V_1|$, $n_2 = |V_2|$, $n_3 = |V_3|$, $n = \max\{n_1, n_2, n_2\}$ and $\alpha \equiv \max n_i/n_j$.

Now, for the choice of $\delta = n^{-5}$, the bounded degree and discrepancy properties in (11), (12), and (13) hold for random tripartite hypergraphs. This finishes the proof of Theorem 2.1.

A.3 Proof of the bounded degree and discrepancy properties in Lemma A.3

We first prove the bounded degree properties of (11) hold with probability at least $1 - \delta$. Applying standard concentration inequality, e.g. Bernstein inequality, we get that for some positive constant $\delta > 0$,

$$\mathbb{P}\left(\deg_1(i) \leq 2pn_2n_3 + \frac{8}{3}\log\frac{3n_1}{\delta}\right) \leq \exp\left(-\frac{(1/2)(pn_2n_3 + (8/3)\log(3n_1/\delta))^2}{(1/3)(pn_2n_3 + (8/3)\log(3n_1/\delta)) + n_2n_3p(1-p)}\right)$$

$$\leq e^{-\log(3n_1/\delta)},$$

for n sufficiently large, and taking union bound over all choices of i, j and k, $\deg_1(i)$, $\deg_2(j)$, and $\deg_3(k)$'s are uniformly bounded with probability at least $1 - \delta/2$.

Similarly, we can apply concentration inequality to bound for some positive constant $\delta > 0$

$$\mathbb{P}\left(\deg_{12}(i,j) \leq 2pn_3 + \frac{8}{3}\log\frac{3n_1n_2}{\delta}\right) \leq \exp\left(-\frac{(1/2)(pn_3 + (8/3)\log(3n_1n_2/\delta))^2}{(1/3)(pn_3 + (8/3)\log(3n_1n_2/\delta)) + n_3p(1-p)}\right)$$

$$\leq e^{-\log(3n_1n_2/\delta)}.$$

16

Applying the union bound over all choices of (i, j), (i, k) and (j, k), we get that the bound holds uniformly with probability at least $1 - \delta/2$.

Next, we prove that the random hyper graphs satisfy the discrepancy properties of (12) and (13). For any given subsets $A_1 \subseteq [n_1]$, $A_2 \subseteq [n_2]$, and $A_2 \subseteq [n_3]$, let a_1, a_2, and a_3 denote the cardinality of the subsets, and $\bar{e}(A_1, A_2, A_3) = p a_1 a_2 a_3$.

Let's assume, without loss of generality, that $a_1 \leq a_2 \leq a_3$. We divide the analysis into two cases depending on the size of the smallest subset. When at least two of the subsets are large, i.e. $a_2 = \Omega(n)$ and $a_3 = \Omega(n)$, then by bounded degree property, we can prove that (12) holds. However, when a_1 and a_2 are small, e.g. $O(1)$, then the first discrepancy no longer holds, and we need a different technique to show concentration.

Case 1. When $a_1 \geq n_1/e$.

From the bounded degree property, we know that $\deg_1(i) \leq 2p n_2 n_3 + (8/3) \ln(3 n_1/\delta)$. Then,

$$
\begin{aligned}
e(A_1, A_2, A_3) &\leq a_1 (2p n_2 n_3 + (8/3) \ln(3 n_1/\delta)) \\
&\leq a_1 (5 p n_2 n_3 + (8/3) \ln(1/\delta)) \\
&\leq 5 a_1 p n_2 n_3 \left(1 + \frac{\ln(1/\delta)}{p n_2 n_3} \right) \\
&\leq 5 e^2 \alpha^2 \bar{e}(A_1, A_2, A_3) \left(1 + \frac{\ln(1/\delta)}{p n_2 n_3} \right) .
\end{aligned}
$$

Case 2. When $a_1 < n_1/e$.

We use the following bound on sum of indicator variables deviating from the mean :

$$
\mathbb{P}\big(e(A_1, A_2, A_3) \geq t \bar{e}(A_1, A_2, A_3) \big) \leq e^{-(1/3)\, \bar{e}\, t \ln t} , \tag{15}
$$

where we denote $\bar{e}(A_1, A_2, A_3)$ by $\bar{e}$, which holds for $t \geq 4$. For the bound holds with probability at least $1 - \delta$, we require

$$
e^{-(1/3)\bar{e} t \ln t} \binom{n_1}{a_1} \binom{n_2}{a_2} \binom{n_3}{a_3} \leq \frac{\delta}{n_1 n_2 n_3} ,
$$

where the term $1/(n_1 n_2 n_3)$ is chosen to compensate for the union bound over all choices of a_1, a_2 and a_3. Simplifying the combinatorial terms, we get

$$
e^{-(1/3)\bar{e} t \ln t} \left(\frac{e n_1}{a_1} \right)^{a_1} \left(\frac{e n_2}{a_2} \right)^{a_2} \left(\frac{e n_3}{a_3} \right)^{a_3} e^{\ln(n_1 n_2 n_3/\delta)} \leq 1 .
$$

Equivalently,

$$
a_1 \ln \left(\frac{e n_1}{a_1} \right) + a_2 \ln \left(\frac{e n_2}{a_2} \right) + a_3 \ln \left(\frac{e n_3}{a_3} \right) + \ln \left(\frac{n_1 n_2 n_3}{\delta} \right) \leq \frac{\bar{e} t \ln t}{3} .
$$

We assumed that $a_1 \leq n_1/e$, and since $x \ln(n_1/x)$ is monotone in $x \in [1, n_1/e]$, we know that $a_1 \ln(e n_1/a_1) \geq \ln n_1$.

$$
4 a_1 \ln \left(\frac{e n_1}{a_1} \right) + a_2 \ln \left(\frac{e n_2}{a_2} \right) + a_3 \ln \left(\frac{e n_3}{a_3} \right) + \ln \left(\frac{\alpha^2}{\delta} \right) \leq \frac{\bar{e} t \ln t}{3} .
$$

To lighten the notations, let's suppose $a_1 \ln(e\, n_1/a_1) \leq a_2 \ln(e\, n_2/a_2) \leq a_3 \ln(e\, n_3/a_3)$. Let t' be the smallest number such that $(3/\bar{e})\big(6 a_3 \ln(e\, n_3/a_3) + \ln(\alpha^2/\delta) \big) = t' \ln t'$.

17

For the regime of parameters such that $t' \le 4$, then $e(A_1, A_2, A_3) \le 4\bar{e}(A_1, A_2, A_3)$ with probability at least $1 - \delta$ the bounded discrepancy condition, in particular the first one, holds.

For the regime of parameters such that $t' > 4$, we can apply (15) to get that with probability at least $1 - \delta$, the following holds uniformly for all choices of A_1, A_2, and A_3:

$$e(A_1, A_2, A_3) \le t'\bar{e}(A_1, A_2, A_3) .$$

Since we defined t' to satisfy $\bar{e}t' \ln t' = 18a_3 \ln(e\, n_3/a_3) + 3\ln(\alpha^2/\delta)$, we have

$$e(A_1, A_2, A_3) \ln t' \le 18a_3 \ln(e\, n_3/a_3) + 3\ln(\alpha^2/\delta) .$$

As t' upper bounds $e(A_1, A_2, A_3)/\bar{e}$, we have

$$e(A_1, A_2, A_3) \ln\left(\frac{e(A_1, A_2, A_3)}{\bar{e}(A_1, A_2, A_3)}\right) \le 18a_3 \ln(e\, n_3/a_3) + 3\ln(\alpha^2/\delta) .$$

A.4 Proof of Thresholding

Lemma A.4. *Let $\mathbf{u}_\ell, 1 \le \ell \le r$ be such that $\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\|_2 \le \alpha$ where $\alpha < 1/4$. Also, let $\mathbf{u}_\ell^*, 1 \le \ell \le r$ be μ-incoherent unit vectors. Now define $\widetilde{\mathbf{u}_\ell}$ as:*

$$\widetilde{\mathbf{u}_\ell}(i) = \begin{cases} \mathbf{u}_\ell(i) & \text{if } |\mathbf{u}_\ell(i)| \le \frac{\mu}{\sqrt{n}}, \\ sign(\mathbf{u}_\ell(i))\frac{\mu}{\sqrt{n}} & \text{if } |\mathbf{u}_\ell(i)| > \frac{\mu}{\sqrt{n}}. \end{cases}$$

Also, let $\widehat{\mathbf{u}_\ell} = \widetilde{\mathbf{u}_\ell}/\|\widetilde{\mathbf{u}_\ell}\|_2$. Then, $\|\widehat{\mathbf{u}_\ell} - \mathbf{u}_\ell^*\|_2 \le 3\alpha \ \forall 1 \le \ell \le r$ and each $\widehat{\mathbf{u}_\ell}$ is 2μ-incoherent.

Proof. As $\|\mathbf{u}_\ell^*\|_\infty \le \frac{\mu}{\sqrt{n}}$, hence $\|\widetilde{\mathbf{u}_\ell} - \mathbf{u}_\ell^*\|_2 \le \|\mathbf{u}_\ell - \mathbf{u}_\ell^*\|_2 \le \alpha$, $\forall \ell$. This also implies that $1 - \alpha \le \|\widetilde{\mathbf{u}_\ell}\|_2 \le 1$. Hence,

$$\|\widehat{\mathbf{u}_\ell} - \mathbf{u}_\ell^*\|_2 \le \|\widetilde{\mathbf{u}_\ell} - \mathbf{u}_\ell^*\|_2 + \left(\frac{1}{\|\widetilde{\mathbf{u}_\ell}\|_2} - 1\right) \le 3\alpha.$$

Moreover, $\|\widehat{\mathbf{u}_\ell}\|_\infty \le \frac{\mu}{\sqrt{n}\cdot(1-\alpha)} \le \frac{2\mu}{\sqrt{n}}$. Hence proved. $\qquad\square$

B Alternating Minimization Analysis

B.1 Main theorem for rank-two analysis

In this section, we provide convergence analysis for Algorithm 1 for the special case of a rank-2 orthonormal tensor T with equal singular values, i.e. $T = \mathbf{u}_1^* \otimes \mathbf{u}_1^* \otimes \mathbf{u}_1^* + \mathbf{u}_2^* \otimes \mathbf{u}_2^* \otimes \mathbf{u}_2^*$, where $\mathbf{u}_1^*, \mathbf{u}_2^* \in \mathbb{R}^n$ are orthonormal vectors satisfying μ-incoherence, i.e., $\|\mathbf{u}_i^*\|_\infty \le \mu/\sqrt{n}$. The purpose of this example is to highlight the proof ideas and we fix σ_1, σ_2 to be both one at each step of Algorithm 1 for simplicity. The following theorem proves the desired linear convergence. Let $[\mathbf{u}_1^t, \mathbf{u}_2^t]$ denote the current estimate at the t-th iteration of Algorithm 1. For brevity, we drop the superscript indexing time and let $[\mathbf{u}_1, \mathbf{u}_2]$ denote $[\mathbf{u}_1^t, \mathbf{u}_2^t]$ whenever it is clear from the context.

Theorem B.1. *If $\mathbf{u}_1$ and $\mathbf{u}_2$ are 2μ-incoherent, then there exists a positive constant C such that for $p \ge C\frac{\mu^3 \log^2 n}{n^{1.5}}$ the following holds (w.p. $\ge 1 - \log(1/\epsilon)/n^8$):*

$$d_\infty\left([\mathbf{u}_1^{t+1}, \mathbf{u}_2^{t+1}], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right) \le \frac{1}{4}d_\infty\left([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right),$$

where $d_\infty([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]) = \max_{i, 1 \le i \le 2} \|\mathbf{u}_i - \mathbf{u}_i^*\|_2$. Moreover, $\mathbf{u}_1^{t+1}$, $\mathbf{u}_2^{t+1}$ are both 2μ-incoherent.

Proof. We claim that with probability at least $1 - 1/n^8$,

$$\|\mathbf{u}_i^{t+1} - \mathbf{u}_i^*\|_2 \;\leq\; \frac{1}{4} \cdot d_\infty\left([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right) \;,$$

for both $i \in \{1, 2\}$. This proves the desired bound. Incoherence of $[\mathbf{u}_1^{t+1}, \mathbf{u}_2^{t+1}]$ follows from Lemma B.2. Without loss of generality, we only prove the claim for $i = 1$. Recall that $\widehat{\mathbf{u}}_1^{t+1}$ is the solution of the least squares problem in Step 11 of Algorithm 1, and can be written as

$$\widehat{\mathbf{u}}_1^{t+1}(i) = \frac{\sum_{jk} \delta_{ijk} \mathbf{u}_1(j) \mathbf{u}_1(k) \mathbf{u}_1^*(j) \mathbf{u}_1^*(k)}{\sum_{jk} \delta_{ijk} (\mathbf{u}_1(j))^2 (\mathbf{u}_1(k))^2} \mathbf{u}_1^*(i) + \frac{\sum_{jk} \delta_{ijk} \mathbf{u}_1(j) \mathbf{u}_1(k) (\mathbf{u}_2^*(i) \mathbf{u}_2^*(j) \mathbf{u}_2^*(k) - \mathbf{u}_2(i) \mathbf{u}_2(j) \mathbf{u}_2(k))}{\sum_{jk} \delta_{ijk} (\mathbf{u}_1(j))^2 (\mathbf{u}_1(k))^2}. \tag{16}$$

Note that the update that can be written in a vector form:

$$\begin{aligned}
\widehat{\mathbf{u}}^{t+1} = \langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 \mathbf{u}_1^* + \langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 \mathbf{u}_2^* - \langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 \mathbf{u}_2 + B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 B - C) \mathbf{u}_1^* \\
+ B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 B - F) \mathbf{u}_2^* - B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 B - G) \mathbf{u}_2,
\end{aligned} \tag{17}$$

where B, C, F, G are all diagonal matrices, s.t.,

$$B_{ii} = \sum_{jk} \delta_{ijk} (\mathbf{u}_1(j))^2 (\mathbf{u}_1(k))^2, \quad C_{ii} = \sum_{jk} \delta_{ijk} \mathbf{u}_1(j) \mathbf{u}_1(k) \mathbf{u}_1^*(j) \mathbf{u}_1^*(k),$$

$$F_{ii} = \sum_{jk} \delta_{ijk} \mathbf{u}_1(j) \mathbf{u}_1(k) \mathbf{u}_2^*(j) \mathbf{u}_2^*(k), \quad G_{ii} = \sum_{jk} \delta_{ijk} \mathbf{u}_1(j) \mathbf{u}_1(k) \mathbf{u}_2(j) \mathbf{u}_2(k) \;. \tag{18}$$

Let $\widehat{\mathbf{u}}_1^{t+1} - \langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 \mathbf{u}_1^* = \mathrm{err}^0 + \mathrm{err}^1 + \mathrm{err}^2$, such that

$$\begin{aligned}
\mathrm{err}^0 &= \langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 \mathbf{u}_2^* - \langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 \mathbf{u}_2, \\
\mathrm{err}^1 &= B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 B - C) \mathbf{u}^*, \\
\mathrm{err}^2 &= B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 B - F) \mathbf{u}_2^* - B^{-1}(\langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 B - G) \mathbf{u}_2 \;.
\end{aligned} \tag{19}$$

We separate the analysis for each of the error terms. Using Lemma B.3, we have:

$$\|\mathrm{err}^0\|_2 \leq 4 d_\infty\left([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right) \|\mathbf{u}_2 - \mathbf{u}_2^*\|_2. \tag{20}$$

Setting $p \geq C \frac{\mu^3 \log^2 n}{\gamma^2 n^{3/2}}$ for a γ to be chosen appropriately later and using Lemma B.7 and Lemma B.5, we have (w.p. $\geq 1 - 2/n^9$):

$$\|\mathrm{err}^1\|_2 \leq \frac{\gamma}{1 - \gamma} \|\mathbf{u}_1 - \mathbf{u}_1^*\|_2. \tag{21}$$

Similarly, using Lemma B.4 and $p \geq C \frac{\mu^3 \log^2 n}{\gamma^2 n^{3/2}}$, we have (w.p. $\geq 1 - 1/n^9$):

$$\|\mathrm{err}^2\| \leq 8 \frac{\gamma}{1 - \gamma} \cdot \|\mathbf{u}_2 - \mathbf{u}_2^*\|_2. \tag{22}$$

We want to upper bound the error:

$$\|\widehat{\mathbf{u}}_1^{t+1} - \mathbf{u}_1^*\|_2 \;\leq\; \|\widehat{\mathbf{u}}_1^{t+1} - \langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 \mathbf{u}_1^*\|_2 + \|(\langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 - 1)\mathbf{u}_1^*\|_2 \;.$$

Since $1 - \langle \mathbf{u}_1, \mathbf{u}_1^* \rangle^2 = (1/2)\|\mathbf{u}_1 - \mathbf{u}_1^*\|_2^2$, we have from (20), (21), and (22) that (w.p. $\geq 1 - 10/n^9$):

$$\|\widehat{\mathbf{u}}_1^{t+1} - \mathbf{u}_1^*\|_2 \;\leq\; \left(\frac{\gamma}{1 - \gamma} + \frac{\|\mathbf{u}_1 - \mathbf{u}_1^*\|_2}{2}\right) \|\mathbf{u}_1 - \mathbf{u}_1^*\|_2 + \left(8 \frac{\gamma}{1 - \gamma} + 4 d_\infty\left([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right)\right) \|\mathbf{u}_2 - \mathbf{u}_2^*\|_2 \;.$$

Setting $\gamma \leq 1/200$ and for $d_\infty\left([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*]\right) \leq 1/200$ as per our assumption, this proves the desired bound. $\qquad\square$

B.2 Technical lemmas for rank-two analysis

The next lemma shows that all our estimates are 2μ-incoherent, which in turn allows us to bound the error in the above proof effectively. Note that the incoherence of the updates do not increase beyond a global constant (2μ). Let $\widehat{\mathbf{u}}_1^{t+1}$ be obtained by update (16) and let $\mathbf{u}_1^{t+1} = \widehat{\mathbf{u}}_1^{t+1}/\|\widehat{\mathbf{u}}_1^{t+1}\|_2$.

Lemma B.2. *Under the hypotheses of Theorem B.1, $\mathbf{u}_1^{t+1}$ is 2μ-incoherent with probability at least $1-1/n^9$.*

Proof. Using (16) and the definitions of B, C, F, G given in (18), we have:

$$|\widehat{\mathbf{u}}_1^{t+1}(i)| \leq \frac{|C_{ii}|}{|B_{ii}|}\frac{\mu}{\sqrt{n}} + \frac{|F_{ii}|}{|B_{ii}|}\frac{\mu}{\sqrt{n}} + \frac{|G_{ii}|}{|B_{ii}|}\frac{2\mu}{\sqrt{n}} \;\leq\; \frac{2\mu}{\sqrt{n}}\;, \tag{23}$$

where the second inequality follows by bounds on $B_{ii}, C_{ii}, F_{ii}, G_{ii}$ obtained using Lemma B.5 and the distance bound $d_\infty([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*])$. $\qquad\square$

Next, we bound the first error term in (17).

Lemma B.3. *Let $\mathbf{u} = \mathbf{u}^* + \mathbf{d}^u$ and $\mathbf{v} = \mathbf{v}^* + \mathbf{d}^v$, where $\mathbf{u}, \mathbf{u}^*, \mathbf{v}, \mathbf{v}^*$ are all unit vectors and $\mathbf{u}^* \perp \mathbf{v}^*$. Also, let $\|\mathbf{d}^u\|_2 \leq 1$ and $\|\mathbf{d}^v\|_2 \leq 1$. Then, the following holds:*

$$\|\langle \mathbf{u}, \mathbf{v}^*\rangle^2 \mathbf{v}^* - \langle \mathbf{u}, \mathbf{v}\rangle^2 \mathbf{v}\| \leq 4(\|\mathbf{d}^u\|_2 + \|\mathbf{d}^v\|_2)\|\mathbf{d}^v\|_2.$$

Proof. Note that,

$$\langle \mathbf{u}, \mathbf{v}\rangle^2 = (\langle \mathbf{u}, \mathbf{v}^*\rangle + \langle \mathbf{u}, \mathbf{d}^v\rangle)^2 = \langle \mathbf{u}, \mathbf{v}^*\rangle^2 + \langle \mathbf{u}, \mathbf{d}^v\rangle^2 + 2\langle \mathbf{u}, \mathbf{v}^*\rangle\langle \mathbf{u}, \mathbf{d}^v\rangle. \tag{24}$$

Hence,

$$\|\langle \mathbf{u}, \mathbf{v}^*\rangle^2 \mathbf{v}^* - \langle \mathbf{u}, \mathbf{v}\rangle^2 \mathbf{v}\|_2 = \|\langle \mathbf{u}, \mathbf{v}\rangle^2 \mathbf{d}^v - \langle \mathbf{u}, \mathbf{d}^v\rangle^2 \mathbf{v}^* - 2\langle \mathbf{u}, \mathbf{v}^*\rangle\langle \mathbf{u}, \mathbf{d}^v\rangle\mathbf{v}^*\|_2,$$

$$\leq \langle \mathbf{u}, \mathbf{v}\rangle^2 \|\mathbf{d}^v\|_2 + \|\mathbf{d}^v\|^2 + 2|\langle \mathbf{u}, \mathbf{v}^*\rangle|\|\mathbf{d}^v\|_2. \tag{25}$$

Now, $\langle \mathbf{u}, \mathbf{v}^*\rangle = \langle \mathbf{d}^u, \mathbf{v}^*\rangle \leq \|\mathbf{d}^u\|_2$. Also, $\langle \mathbf{u}, \mathbf{v}\rangle^2 \leq \langle \mathbf{u}, \mathbf{v}\rangle = (\langle \mathbf{d}^u, \mathbf{v}^*\rangle + \langle \mathbf{u}^*, \mathbf{d}^v\rangle + \langle \mathbf{d}^u, \mathbf{d}^v\rangle) \leq 2(\|\mathbf{d}^u\| + \|\mathbf{d}^v\|)$. Lemma now follows by combining the above observations with (25). $\qquad\square$

Now, we bound the third error term in (17). Note that although the two individual terms $((\langle \mathbf{u}_1, \mathbf{u}_2^*\rangle^2 B - F)\mathbf{u}_2^*$ and $(\langle \mathbf{u}_1, \mathbf{u}_2\rangle^2 B - G)\mathbf{v})$ are both small, still it is critical to bound the difference as the individual terms can be as large as a constant, even when $\mathbf{u}_1 = \mathbf{u}_1^*$ and $\mathbf{u}_2 = \mathbf{u}_2^*$. However, the difference goes down linearly with $\|\mathbf{u}_2 - \mathbf{u}_2^*\|_2$.

Lemma B.4. *Let B, C, F, G be defined as in (18). Also, let the assumptions of Theorem B.1 hold. Also, let $p \geq C\frac{\mu^3 \log^2 n}{\gamma^2 n^{3/2}}$, where $C > 0$ is a global constant. Then, the following holds with probability $\geq 1 - 4/n^9$:*

$$\|(\langle \mathbf{u}_1, \mathbf{u}_2^*\rangle^2 B - F)\mathbf{u}_2^* - (\langle \mathbf{u}_1, \mathbf{u}_2\rangle^2 B - G)\mathbf{u}_2\|_2 \leq 8\gamma\|\mathbf{u}_2 - \mathbf{u}_2^*\|_2.$$

Proof. Let $\mathbf{u}_2 = \mathbf{u}_2^* + \mathbf{d}_2^u$ and $\mathbf{u}_1 = \mathbf{u}_1^* + \mathbf{d}_1^u$. Then,

$$(\langle \mathbf{u}_1, \mathbf{u}_2\rangle^2 B - G)\mathbf{u}_2 = (\langle \mathbf{u}_1, \mathbf{u}_2\rangle^2 B - G)\mathbf{u}_2^* + (\langle \mathbf{u}_1, \mathbf{u}_2\rangle^2 B - G)\mathbf{d}_2^u. \tag{26}$$

Now,

$$G_{ii} = \sum_{jk} \delta_{ijk}\mathbf{u}_1(j)\mathbf{u}_1(k)\mathbf{u}_2(j)\mathbf{u}_2(k) = \sum_{jk} \delta_{ijk}\mathbf{u}_1(j)\mathbf{u}_1(k)(\mathbf{u}_2^*(j) + \mathbf{d}_2^u(j))(\mathbf{u}_2^*(k) + \mathbf{d}_2^u(k))$$

$$= \sum_{jk} \delta_{ijk}\mathbf{u}_1(j)\mathbf{u}_1(k)\left(\mathbf{u}_2^*(j)\mathbf{u}_2^*(k) + \mathbf{d}_2^u(j)\mathbf{u}_2^*(k) + \mathbf{u}_2^*(j)\mathbf{d}_2^u(k) + \mathbf{d}_2^u(j)\mathbf{d}_2^u(k)\right) = F_{ii} + D_{ii}^1 + D_{ii}^2 + D_{ii}^3.$$

$$\tag{27}$$

Hence, using (24), and (27), we have:

$$(\langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 B - G) = (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 B - F) + (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle \langle \mathbf{u}_1, \mathbf{d}_2^u \rangle B - D^1)$$
$$+ (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle \langle \mathbf{u}_1, \mathbf{d}_2^u \rangle B - D^2) + (\langle \mathbf{u}_1, \mathbf{d}_2^u \rangle^2 B - D^3) \quad (28)$$

Combining the above equation with (26), we get:

$$(\langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 B - G)\mathbf{u}_2 - (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle^2 B - F)\mathbf{u}_2^* = (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle \langle \mathbf{u}_1, \mathbf{d}_2^u \rangle B - D^1)\mathbf{u}_2^* + (\langle \mathbf{u}_1, \mathbf{u}_2^* \rangle \langle \mathbf{u}_1, \mathbf{d}_2^u \rangle B - D^2)\mathbf{u}_2^*$$
$$+ (\langle \mathbf{u}_1, \mathbf{d}_2^u \rangle^2 B - D^3)\mathbf{u}_2^* - (\langle \mathbf{u}_1, \mathbf{u}_2 \rangle^2 B - G)\mathbf{d}_2^u. \quad (29)$$

Lemma now follows using Lemma B.7, B.8, and the above equation. $\qquad \square$

We now present a few technical lemmas that are critical to our proofs of the above given lemmas.

Lemma B.5. *Let* $\mathbf{u}, \mathbf{u}^* \in \mathbb{R}^n$ *be* μ*-incoherent unit vectors. Also, let* $\delta_{jk}, 1 \leq j \leq n, 1 \leq k \leq n$ *be i.i.d. Bernoulli random variables with* $\delta_{jk} = 1$ *w.p.* $p \geq C\mu^4 \log^3 n/(\gamma^2 \cdot n^2)$.

Then, the following holds with probability $\geq 1 - 1/n^{10}$:

$$\left| \frac{1}{p} \sum_{jk} \delta_{jk} \mathbf{u}(j)\mathbf{u}^*(j)\mathbf{u}(k)\mathbf{u}^*(k) - \langle \mathbf{u}, \mathbf{u}^* \rangle^2 \right| \leq \gamma.$$

where $\gamma \leq C/\log n$*, where* $C > 0$ *is a global constant.*

Lemma B.6. *Let* $\mathbf{u} \in \mathbb{R}^n$ *be* μ*-incoherent unit vectors. Also, let* $a, b \in \mathbb{R}^n$ *be s.t.* $|a_i| \leq \frac{\mu}{\sqrt{n}}$ *and* $\|a\|_2 \leq 1$.

Also, let $\delta_{jk}, 1 \leq j \leq n, 1 \leq k \leq n$ *be i.i.d. Bernoulli random variables with* $\delta_{jk} = 1$ *w.p.* $p \geq \frac{C\mu^3}{\gamma^2 n^{1.5}}$.

Then, the following holds with probability $\geq 1 - 1/n^{10}$:

$$\left| \frac{1}{p} \sum_{jk} \delta_{jk} \mathbf{u}(j)a(j)\mathbf{u}(k)b(k) - \langle \mathbf{u}, a \rangle \langle \mathbf{u}, b \rangle \right| \leq \gamma \|b\|_2.$$

where $\gamma \leq C/\log n$*, where* $C > 0$ *is a global constant.*

Lemma B.7. *Let* $\mathbf{u}$ *be a fixed unit vector and let* a, b, c *be fixed vectors in* $\mathbb{R}^n$*. Also, let all* $\mathbf{u}, a, b, c \in \mathbb{R}^n$ *be s.t. their* L_∞ *norm is bounded by* $\frac{\mu}{\sqrt{n}}$ *and* L_2 *norm is bounded by 1. Also, let* $p \geq \frac{C\mu^3(\log^2 n)}{\gamma^2 \cdot n^{3/2}}$*, where* $C > 0$ *is a global constant. Then the following holds (w.p.* $\geq 1 - 2/n^{10}$*):*

$$\|(\langle \mathbf{u}, a \rangle \langle \mathbf{u}, b \rangle B - R)c\|_2 \leq \gamma \sqrt{1 - \langle \mathbf{u}, a \rangle^2 \langle \mathbf{u}, b \rangle^2},$$

where B, R *are both diagonal matrices with* $B(i,i) = \frac{1}{p} \sum_{jk} \delta_{ijk}(\mathbf{u}(j))^2(\mathbf{u}(k))^2$*, and* $R(i,i) = \frac{1}{p} \sum_{jk} \delta_{ijk}\mathbf{u}(j)\mathbf{u}(k)a(j)b(k)$.

Lemma B.8. *Let* $\mathbf{u}$ *be a fixed unit vector and let* a, b *be fixed vectors in* $\mathbb{R}^n$*. Also, let all* $\mathbf{u}$ *be* μ*-incoherent unit vectors, and* a *be such that* $\|a\|_\infty \leq \frac{\mu}{\sqrt{n}}$ *and* $\|a\|_2 \leq 1$*. Also, let* $p \geq \frac{C\mu^3(\log^2 n)}{\gamma^2 \cdot n^{3/2}}$*, where* $C > 0$ *is a global constant. Then the following holds (w.p.* $\geq 1 - 2/n^{10}$*):*

$$\|(\langle \mathbf{u}, a \rangle \langle \mathbf{u}, b \rangle B - R)\|_2 \leq 2\gamma \|b\|_2$$

where B, R *are diagonal matrices s.t.* $B(i,i) = \frac{1}{p} \sum_{jk} \delta_{ijk}(\mathbf{u}(j))^2(\mathbf{u}(k))^2$*,* $R(i,i) = \frac{1}{p} \sum_{jk} \delta_{ijk}\mathbf{u}(j)\mathbf{u}(k)a(j)b(k)$.

B.3   Proofs of Technical Lemmas

Proof of Lemma B.5. Let $X_{jk} = \frac{1}{p}\delta_{jk}\mathbf{u}(j)\mathbf{u}^*(j)\mathbf{u}(k)\mathbf{u}^*(k)$. Note that, $|X_{jk}| \leq \frac{\mu^4}{pn^2}$. Also,

$$\mathbb{E}[\sum_{jk} X_{jk}^2] = \frac{1}{p} \sum_{jk} (\mathbf{u}(j))^2(\mathbf{u}^*(j))^2(\mathbf{u}(k))^2(\mathbf{u}^*(k))^2 \leq \frac{\mu^4}{pn^2}.$$

21

Hence, using Bernstein's inequality, we have:

$$Pr(|\sum_{jk} X_{jk} - \mathbb{E}[\sum_{jk} X_{jk}]| > t) \leq \exp(-\frac{pn^2}{\mu^4} \cdot \frac{t^2/2}{1+t/3}).$$

Lemma now follows by selecting $t = C/\log n$. $\square$

Proof of Lemma B.6. Let $X_{jk} = \frac{1}{p}\delta_{jk}\mathbf{u}(j)a(j)\mathbf{u}(k)b(k)$. Note that, $|X_{jk}| \leq \frac{\mu^3\|b\|_2}{pn^{1.5}}$. Also,

$$\mathbb{E}[\sum_{jk} X_{jk}^2] = \frac{1}{p}\sum_{jk}(\mathbf{u}(j))^2(a(j))^2(\mathbf{u}(k))^2(b(k))^2 \leq \frac{\mu^4\|b\|^2}{pn^2} \leq \frac{\mu^3\|b\|^2}{pn^{1.5}}.$$

Hence, using Bernstein's inequality, we have:

$$Pr(|\sum_{jk} X_{jk} - \mathbb{E}[\sum_{jk} X_{jk}]| > t) \leq \exp(-\frac{pn^{1.5}}{\mu^3} \cdot \frac{t^2/2}{\|b\|_2^2 + \|b\|_2 t/3}).$$

Lemma now follows by selecting $t = \gamma\|b\|_2$. $\square$

Proof of Lemma B.7.

$$(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B - R)c = \frac{1}{p}\sum_{ijk}\delta_{ijk}c_i(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle(u(j))^2(u(k))^2 - u(j)u(k)a(j)b(k))\mathbf{e}_i = \sum_{ijk} Z_{ijk}, \qquad (30)$$

where $Z_{ijk} = \frac{1}{p}\delta_{ijk}c_i(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle(u(j))^2(u(k))^2 - u(j)u(k)a(j)b(k))\mathbf{e}_i$. Note that,

$$\|Z_{ijk} - \mathbb{E}[Z_{ijk}]\|_2 \leq \frac{2}{p}c_i u(j)u(k)\sqrt{1 - \langle \mathbf{u}, a\rangle^2\langle \mathbf{u}, b\rangle^2} \leq \gamma\sqrt{1 - \langle \mathbf{u}, a\rangle^2\langle \mathbf{u}, b\rangle^2},$$

as $p \geq \frac{C\mu^3(\log^2 n)}{\gamma \cdot n^{3/2}}$. Also,

$$\|\sum_{ijk}\mathbb{E}[Z_{ijk}^T Z_{ijk}]\|_2 = \|\frac{1}{p}\sum_{ijk}c_i^2(u(j))^2(u(k))^2(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle u(j)u(k) - a(j)b(k))^2\|_2 \leq \frac{1}{p}\frac{\mu^4}{n^2}(1 - \langle \mathbf{u}, a\rangle^2\langle \mathbf{u}, b\rangle^2).$$

Hence, for p and γ mentioned above, we have:

$$\|\sum_{ijk}\mathbb{E}[Z_{ijk}^T Z_{ijk}]\|_2 \leq \gamma(1 - \langle \mathbf{u}, a\rangle^2\langle \mathbf{u}, b\rangle^2).$$

Lemma now follows by using Bernstein's inequality and the fact that $\sum_{ijk} Z_{ijk} = 0$. $\square$

Proof of Lemma B.8. Consider the i-th element of the diagonal matrix $(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B - R) = \langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B(i,i) - R(i,i)$. Now, using Lemma B.5, $|B(i,i)| \leq 1 + \gamma$ w.p. $\geq 1 - 1/n^{10}$. Similarly, using Lemma B.6, $|R(i,i) - \langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle| \leq \gamma\|b\|_2$. Hence, w.p. $\geq 1 - 1/n^{10}$, we have:

$$|\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B(i,i) - R(i,i)| \leq 2\gamma\|b\|_2.$$

Lemma now follows by observing that $\|(\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B - R)\|_2 = \max_i |\langle \mathbf{u}, a\rangle\langle \mathbf{u}, b\rangle B(i,i) - R(i,i)|$ and using the above mentioned bound with union bound. $\square$

B.4 Proof of Theorem 2.3 and general rank-r analysis of alternating minimization

Proof. We prove the theorem by showing the following for all q:

$$\sigma_q^* \left((\Delta_q^\sigma)^{t+1} + \|\mathbf{d}_q^{t+1}\|_2 \right) \leq \frac{1}{2} d_\infty([U, \Sigma], [U^*, \Sigma^*]).$$

The update for $\widehat{\mathbf{u}}_q^{t+1}$ is given by:

$$\widehat{\mathbf{u}}_q^{t+1}(i) = \frac{\sum_{jk} \delta_{ijk} \sigma_q^* \cdot \mathbf{u}_q(j) \mathbf{u}_q(k) \mathbf{u}_q^*(j) \mathbf{u}_q^*(k)}{\sum_{jk} \delta_{ijk} \mathbf{u}_q(j)^2 \mathbf{u}_q(k)^2} \mathbf{u}_q^*(i)$$
$$+ \frac{\sum_{\ell \neq q} \sum_{jk} \delta_{ijk} \mathbf{u}_q(j) \mathbf{u}_q(k) (\sigma_\ell^* \cdot \mathbf{u}_\ell^*(i) \mathbf{u}_\ell^*(j) \mathbf{u}_\ell^*(k) - \sigma_\ell \cdot \mathbf{u}_\ell(i) \mathbf{u}_\ell(j) \mathbf{u}_\ell(k))}{\sum_{jk} \delta_{ijk} \mathbf{u}_q(j)^2 \mathbf{u}_q(k)^2}. \quad (31)$$

It can be written as a vector update,

$$\widehat{\mathbf{u}}_q^{t+1} = \sigma_q^* \langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2 \mathbf{u}_q^* - B^{-1}(\sigma_q^* \langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2 B - \sigma_q^* C) \mathbf{u}_q^* + \sum_{\ell \neq q} \left(\sigma_\ell^* \langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \sigma_\ell \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell \right)$$
$$+ \sum_{\ell \neq q} B^{-1} \left(\sigma_\ell^* \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell) \mathbf{u}_\ell^* - \sigma_\ell \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell) \mathbf{u}_\ell \right), \quad (32)$$

where B, C, F_ℓ, G_ℓ are all diagonal matrices, s.t.,

$$B(i,i) = \sum_{jk} \delta_{jk} \mathbf{u}_q(j)^2 \mathbf{u}_q(k)^2, \ \ C(i,i) = \sum_{jk} \delta_{jk} \mathbf{u}_q(j) \mathbf{u}_q^*(j) \mathbf{u}_q(k) \mathbf{u}_q^*(k),$$

$$F_\ell(i,i) = \sum_{jk} \delta_{ijk} \mathbf{u}_q(j) \mathbf{u}_q(k) \mathbf{u}_\ell^*(j) \mathbf{u}_\ell^*(k), \ \text{ and } \ G_\ell(i,i) = \sum_{jk} \delta_{ijk} \mathbf{u}_q(j) \mathbf{u}_q(k) \mathbf{u}_\ell(j) \mathbf{u}_\ell(k). \quad (33)$$

We decompose the error terms $\widehat{\mathbf{u}}_q^{t+1} - \sigma_q^* \mathbf{u}_q^* = \text{err}_q^0 + \sum_{\ell \neq q} (\text{err}_\ell^1 + \text{err}_\ell^2)$ and provide upper bounds for each, where

$$\text{err}_q^0 \equiv \sigma_q^*(\langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2 - 1) \mathbf{u}_q^* - \sigma_q^* B^{-1}(\langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2 B - C) \mathbf{u}_q^*,$$
$$\text{err}_\ell^1 \equiv \sigma_\ell^* \langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \sigma_\ell \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell,$$
$$\text{err}_\ell^2 \equiv B^{-1} \left(\sigma_\ell^* \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell) \mathbf{u}_\ell^* - \sigma_\ell \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell) \mathbf{u}_\ell \right). \quad (34)$$

Using Lemma B.7, we have for all p satisfying $p \geq (C\mu^3 (\log n)^2)/(\gamma^2 n^{3/2})$, with probability at least $1 - 2/n^{10}$:

$$\|\text{err}_q^0\|_2 \leq \sigma_q^* \left(\sqrt{1 - \langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2} + 2\gamma \right) \sqrt{1 - \langle \mathbf{u}_q, \mathbf{u}_q^* \rangle^2} \leq \sigma_q^* (\|\mathbf{d}_q\|_2 + 2\gamma) \|\mathbf{d}_q\|_2. \quad (35)$$

Eventually, we set $\gamma \leq \frac{1}{1600r} \cdot \frac{\sigma_{min}^*}{\sigma_{max}^*}$ to prove the theorem. Using Lemma B.10, we have (w.p. $\geq 1 - 1/n^8$):

$$\sum_{\ell \neq q} \|\text{err}_\ell^1\|_2 \leq 8 \sum_{\ell \neq q} (\|\mathbf{d}_q\|_2 + \|\mathbf{d}_\ell\|_2) \cdot \sigma_\ell^* \cdot (\|\mathbf{d}_\ell\|_2 + \Delta_\ell^\sigma). \quad (36)$$

Using Lemma B.11, we get (w.p. $\geq 1 - 1/n^8$):

$$\sum_{\ell \neq q} \|\text{err}_\ell^2\|_2 \leq 16\gamma \sum_{\ell \neq q} \sigma_\ell^* \cdot (\Delta_\ell^\sigma + \|\mathbf{d}_\ell\|_2). \quad (37)$$

Using (31), (35), (36), (37), we have (w.p. $\geq 1 - 3/n^8$):

$$\widehat{\mathbf{u}}_q^{t+1} = \sigma_q^{t+1} \mathbf{u}_q^{t+1} = \sigma_q^* \mathbf{u}_q^* + \text{err}_q, \quad (38)$$

23

where,

$$\|\text{err}_q\|_2 \ \leq \ \sigma_q^* \left(\|\mathbf{d}_\ell\|_2 + 2\gamma\right) \|\mathbf{d}_q\|_2 + 8 \sum_{\ell \neq q} (\|\mathbf{d}_q\|_2 + \|\mathbf{d}_\ell\|_2 + 2\gamma)\, \sigma_\ell^* \left(\|\mathbf{d}_\ell\|_2 + \Delta_\ell^\sigma\right). \tag{39}$$

Now, since $\|\mathbf{d}_\ell\|_2 \leq \frac{1}{1600r} \cdot \frac{\sigma_{min}^*}{\sigma_{max}^*}, \forall \ell$, and $\gamma \leq \frac{1}{1600r}$, we have (w.p. $\geq 1 - 3/n^8$):

$$\|\text{err}_q\|_2 \ \leq \ \frac{\sigma_q^*}{16} \cdot \frac{\sigma_{min}^*}{\sigma_{max}^*} \|\mathbf{d}_q\|_2 + \frac{1}{16} \cdot \sigma_{min}^* \cdot d_\infty([U,\ \Sigma], [U^*,\ \Sigma^*]), \tag{40}$$

Using (38) and (40), and the fact that $|\sigma_q^{t+1} - \sigma_q^*| \leq |\sigma_q^{t+1} \mathbf{u}_q^{t+1} - \mathbf{u}_q^* \sigma_q^*|$ for normalized vectors $\mathbf{u}_q^{t+1}$ and $\mathbf{u}_q^*$, we have:

$$|\sigma_q^{t+1} - \sigma_q^*| \ \leq \ \frac{\sigma_q^*}{16} \|\mathbf{d}_q\|_2 + \frac{\sigma_{min}^*}{16} d_\infty([U,\ \Sigma], [U^*,\ \Sigma^*]) \leq \frac{\sigma_q^*}{8} d_\infty([U,\ \Sigma], [U^*,\ \Sigma^*]). \tag{41}$$

Similarly, using (38) and (41), we have:

$$\sigma_q^* \|\mathbf{u}_q^{t+1} - \mathbf{u}_q^*\|_2 \ \leq \ \frac{\sigma_q^*}{4} d_\infty([U,\ \Sigma], [U^*,\ \Sigma^*]). \tag{42}$$

That is,

$$(\Delta_q^\sigma)^{t+1} + \|\mathbf{d}_q^{t+1}\|_2 \ \leq \ \frac{1}{2} d_\infty([U,\ \Sigma], [U^*,\ \Sigma^*]). \tag{43}$$

First part of the Theorem now follows by observing that $d_\infty([U^{t+1},\ \Sigma^{t+1}], [U^*,\ \Sigma^*]) = \max_q \sigma_q^* \left((\Delta_q^\sigma)^{t+1} + \|\mathbf{d}_q^{t+1}\|_2\right)$ and by using the above equation.

Second part of the Theorem follows directly from Lemma B.9.

$\square$

B.5 Technical lemmas for general rank-r analysis

Lemma B.9. *Let $\hat{\mathbf{u}}_q^{t+1}$ be obtained by update (31) and let $\mathbf{u}_q^{t+1} = \hat{\mathbf{u}}_q^{t+1}/\|\hat{\mathbf{u}}_q^{t+1}\|_2$. Also, let the conditions given in Theorem 2.3 hold. Then, w.p. $\geq 1 - 1/n^9$, $\mathbf{u}_q^{t+1}$ is 2μ-incoherent.*

Proof. Using (31) and the definitions of B, C, F_ℓ, G_ℓ given in (33), we have:

$$|\hat{\mathbf{u}}_q^{t+1}(i)| \leq \sigma_q^* \frac{|C(i,i)|}{|B(i,i)|} \frac{\mu}{\sqrt{n}} + \sum_\ell \sigma_\ell^* \frac{|F_\ell(i,i)|}{|B(i,i)|} \frac{\mu}{\sqrt{n}} + \sigma_\ell \frac{|G_\ell(i,i)|}{|B(i,i)|} \frac{\mu}{\sqrt{n}},$$

$$\leq \left(\sigma_q^*(1+\gamma)/(1-\gamma) + \sum_\ell \sigma_\ell^*(\gamma + \|\mathbf{d}_\ell\|_2) + 2 \sum_\ell \sigma_\ell^* \cdot (1 + \Delta_\ell^\sigma) \cdot (\gamma + \|\mathbf{d}_\ell\|_2) \right) \mu/\sqrt{n},$$

$$\leq \sigma_q^*(1 + \frac{1}{100}) \cdot \frac{\mu}{\sqrt{n}} \tag{44}$$

where the second inequality follows by bounds on $B_{ii}, C_{ii}, F_{ii}, G_{ii}$ obtained using Lemma B.6 and the distance bound $d_\infty([\mathbf{u}_1, \mathbf{u}_2], [\mathbf{u}_1^*, \mathbf{u}_2^*])$.

Lemma now follows using (44) and the bound on $|\sigma_q^{t+1} - \sigma_q^*|$ given by (43). $\square$

Lemma B.10. *Let $\mathbf{d}_\ell = \mathbf{u}_\ell^* - \mathbf{u}_\ell$ and $\Delta_\ell^\sigma = |\sigma_\ell - \sigma_\ell^*|/\sigma_\ell^*$, where $\|\mathbf{d}_\ell\|_2 \leq 1$. Let $\mathbf{u}_\ell^*, \mathbf{u}_\ell, \forall 1 \leq \ell \leq r$ be unit vectors and let $\langle \mathbf{u}_\ell^*, \mathbf{u}_q^* \rangle = 0, \forall \ell \neq q$. Then, the following holds:*

$$\|\sigma_\ell^* \cdot \langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \sigma_\ell \cdot \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell\| \leq 4\sigma_\ell^*(\|\mathbf{d}_\ell\|_2 + \|\mathbf{d}_q\|_2)(\|\mathbf{d}_\ell\|_2 + \Delta_\ell^\sigma).$$

Proof. Let $\sigma_\ell = \sigma_\ell^* + \Delta_\ell^\sigma$.

Now,

$$\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 = (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle + \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle)^2 = \langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 + \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle^2 + 2\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle. \tag{45}$$

Hence,

$$\begin{aligned}
\|\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell\|_2 &= \|\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{d}_\ell - \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle^2 \mathbf{u}_\ell^* - 2\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle \mathbf{u}_\ell^*\|_2, \\
&\leq \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \|\mathbf{d}_\ell\|_2 + \|\mathbf{d}_\ell\|^2 + 2|\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle| \|\mathbf{d}_\ell\|_2. \tag{46}
\end{aligned}$$

Now, $\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle = \langle \mathbf{d}_q, \mathbf{u}_\ell^* \rangle \leq \|\mathbf{d}_q\|_2$. Also, $\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \leq \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle = (\langle \mathbf{d}_q, \mathbf{u}_\ell^* \rangle + \langle \mathbf{u}_q^*, \mathbf{d}_\ell \rangle + \langle \mathbf{d}_q, \mathbf{d}_\ell \rangle) \leq 2(\|\mathbf{d}_q\| + \|\mathbf{d}_\ell\|)$. Using the above observations with (46), we have:

$$\|\sigma_\ell^* \cdot \langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \sigma_\ell \cdot \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell\| \leq \sigma_\ell^* \|\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 \mathbf{u}_\ell^* - \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 \mathbf{u}_\ell\|_2 + \sigma_\ell^* \cdot \Delta_\ell^\sigma \langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2.$$

Lemma now follows by combining the above equation with the above given bound on $\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2$. $\quad\square$

Lemma B.11. *Let $\mathbf{u}_\ell$, $\mathbf{d}_\ell$, Δ_ℓ^σ, $\forall \ell$ be as defined in Theorem 2.3 and let B, F_ℓ, G_ℓ be as defined in (33). Also, let T and p satisfy assumptions of Theorem 2.3. Then, the following holds with probability $\geq 1 - 4/n^9$:*

$$\|\sigma_\ell^* \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell)\mathbf{u}_\ell^* - \sigma_\ell \cdot (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell\|_2 \leq 8\gamma \cdot \sigma_\ell^* \cdot (\Delta_\ell^\sigma + \|\mathbf{d}_\ell\|_2).$$

Proof.

$$(\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell = (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell^* + (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{d}_\ell. \tag{47}$$

Now,

$$\begin{aligned}
G_\ell(i,i) &= \sum_{jk} \delta_{ijk} \mathbf{u}_q(j)\mathbf{u}_q(k)\mathbf{u}_\ell(j)\mathbf{u}_\ell(k) = \sum_{jk} \delta_{ijk} \mathbf{u}_q(j)\mathbf{u}_q(k)(\mathbf{u}_\ell^*(j) + \mathbf{d}_\ell(j))(\mathbf{u}_\ell^*(k) + \mathbf{d}_\ell(k)) \\
&= \sum_{jk} \delta_{ijk} \mathbf{u}_q(j)\mathbf{u}_q(k)(\mathbf{u}_\ell^*(j)\mathbf{u}_\ell^*(k) + \mathbf{d}_\ell(j)\mathbf{u}_\ell^*(k) + \mathbf{u}_\ell^*(j)\mathbf{d}_\ell(k) + \mathbf{d}_\ell(j)\mathbf{d}_\ell(k)) \\
&= F_\ell(i,i) + D^1(i,i) + D^2(i,i) + D^3(i,i). \tag{48}
\end{aligned}$$

Using (45), and (48), we have:

$$\begin{aligned}
(\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell) = (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell) + (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle B - D^1) \\
+ (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle B - D^2) + (\langle \mathbf{u}_q, \mathbf{d}_\ell \rangle^2 B - D^3) \tag{49}
\end{aligned}$$

Combining the above equation with (47), we get:

$$\begin{aligned}
(\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell - (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell)\mathbf{u}_\ell^* = (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle B - D^1)\mathbf{u}_\ell^* + (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle \langle \mathbf{u}_q, \mathbf{d}_\ell \rangle B - D^2)\mathbf{u}_\ell^* \\
+ (\langle \mathbf{u}_q, \mathbf{d}_\ell \rangle^2 B - D^3)\mathbf{u}_\ell^* - (\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{d}_\ell. \tag{50}
\end{aligned}$$

Hence, using Lemma B.7 and B.8, we get:

$$\sigma_\ell^* \|(\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell - (\langle \mathbf{u}_q, \mathbf{u}_\ell^* \rangle^2 B - F_\ell)\mathbf{u}_\ell^*\|_2 \leq 8\gamma\sigma_\ell^* \|\mathbf{d}_\ell\|_2. \tag{51}$$

Similarly, using Lemma B.7, we have:

$$\Delta_\ell^\sigma \cdot \sigma_\ell^* \|(\langle \mathbf{u}_q, \mathbf{u}_\ell \rangle^2 B - G_\ell)\mathbf{u}_\ell\|_2 \leq \gamma \cdot \Delta_\ell^\sigma \cdot \sigma_\ell^*. \tag{52}$$

Lemma now follows by combining (51), (52), and by using triangular inequality. $\quad\square$

B.6 Proof of Lemma 2.4

Proof.

$$\|\sigma_a(u_a \otimes u_a \otimes u_a) - \sigma_a^*(u_a^* \otimes u_a^* \otimes u_a^*)\|_F$$
$$\leq \ \tilde{\varepsilon}\sigma_a^* + \sigma_a^* \|(u_a \otimes u_a \otimes u_a) - (u_a^* \otimes u_a^* \otimes u_a^*)\|_F$$
$$\leq \ \tilde{\varepsilon}\sigma_a^* + \sigma_a^* \Big(\|(u_a - u_a^*) \otimes u_a^* \otimes u_a^*\|_F + \|u_a \otimes (u_a - u_a^*) \otimes u_a^*\|_F - \|u_a \otimes u_a \otimes (u_a - u_a^*)\|_F \Big)$$
$$\leq \ 4\tilde{\varepsilon}\,\sigma_a^* \,.$$

Similarly, applying Cauchy-Schwartz, we get for $a \neq b$,

$$\langle \sigma_a(u_a \otimes u_a \otimes u_a) - \sigma_a^*(u_a^* \otimes u_a^* \otimes u_a^*), \sigma_b(u_b \otimes u_b \otimes u_b) - \sigma_b^*(u_b^* \otimes u_b^* \otimes u_b^*) \rangle$$
$$\leq \ 16\,\tilde{\varepsilon}^2\,\sigma_a^*\sigma_b^* \,.$$

It follows that

$$\|T - \widehat{T}\|_F^2 \ = \ \sum_{a,b \in [r]} \langle \sigma_a(u_a \otimes u_a \otimes u_a) - \sigma_a^*(u_a^* \otimes u_a^* \otimes u_a^*), \sigma_b(u_b \otimes u_b \otimes u_b) - \sigma_b^*(u_b^* \otimes u_b^* \otimes u_b^*) \rangle$$
$$\leq \ 16\,\tilde{\varepsilon}^2 \Big(\sum_a \sigma_a^*\Big)^2 \ \leq \ (4\,\tilde{\varepsilon}\,r\,\sigma_{\max}^*)^2 \ \leq \ (4\,\tilde{\varepsilon}\,r^{1/2}\,\|T\|_F)^2 \,.$$

□