January 2023

On Partially Observed Tensor Regression

Dinara Miftyakhetdinova
University of Windsor, miftyak@uwindsor.ca

ON PARTIALLY OBSERVED TENSOR REGRESSION

by

Dinara Miftyakhetdinova

A Major Research Paper

Submitted to the Faculty of Graduate Studies

through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Master of Science at the

University of Windsor

Windsor, Ontario, Canada

ON PARTIALLY OBSERVED TENSOR REGRESSION

by

Dinara Miftyakhetdinova

APPROVED BY:

_____

A. Hussein

Department of Mathematics and Statistics

_____

S. Nkurunziza, Advisor

Department of Mathematics and Statistics

December 14, 2022

Author's Declaration of Originality

I hereby certify that I am the sole author of this major paper and that no part of this major paper has been published or submitted for publication.

I certify that, to the best of my knowledge, my major paper does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my major paper, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my major paper and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my major paper, including any final revisions, as approved by my major paper committee and the Graduate Studies office, and that this major paper has not been submitted for a higher degree to any other University or Institution.

Abstract

Tensor data is widely used in modern data science. The interest lies in identifying and characterizing the relationship between tensor datasets and external covariates. These datasets, though, are often incomplete. An efficient nonconvex alternating updating algorithm proposed by J. Zhou et al. in the paper "Partially Observed Dynamic Tensor Response Regression" provides a novel approach. The algorithm handles the problem of unobserved entries by solving an optimization problem of a loss function under the low-rankness, sparsity, and fusion constraints. This analysis aims to understand in detail the proposed algorithms and their theoretical proofs with, potentially, dropping some of the assumptions implied to the model. Also, the efficiency and accuracy of the algorithms on a simulated data and Parkinson's disease real-life dataset will be illustrated.

Acknowledgments

I would first like to thank my supervisor Professor Sévérien Nkurunziza of the Department of Mathematics and Statistics at the University of Windsor. Dr.Nkurunziza supported and advised me from the first day of my Master's program. I greatly appreciate it. He consistently motivated me to work on the paper and steered me in the right direction whenever he thought I needed it.

I would like to acknowledge the professors and students I met during my program. I was able to learn a lot from them, and I am grateful for this opportunity.

Finally, I must express my gratitude to my parents, to my brother, to my partner and to my friends for providing me with constant support and encouragement throughout my study and this paper. This accomplishment would not have been possible without them. Thank you.

Contents

1 Introduction

Tensor data has several potential applications. Compared to the classical regression, tensors show the spacial structure and any correlation among individual voxels. Moreover, with high-dimensional data, converting tensor into vectors/matrices to apply the classic regression models would result in very large parameters. Thus, the interest in regression involving tensors is growing. Tensors are perhaps most advantageous in medical analysis. For example, it has been applied to datasets in biomedical informatics, including MRI scans in studies of Alzheimer's disease (AD) in Thung et al, 2016 [6] and Attention deficit hyperactivity disorder (ADHD) in Zhou et al, 2013 [2]. Tensor data can also be seen in business applications Bruce et al, 2017 [9]. The interest of those studies lie in finding the relationship between given tensor data and external covariates. These datasets, though, are often incomplete in real applications. To solve this issue, some have simply filled out the missing data by using the mean of the data they have or have simply used a smaller sample size. These approaches are limited regarding biomedical data. Thus, it is important to consider other alternatives for dealing with incomplete data.

There are studies that address this problem by completing the tensor data Jain et al, 2014 [10], Xia at el, 2019 [11]. To complete the tensor, some tensor low-rankness and sparsity structures are employed, and unsupervised learning methods are used. Zhou et al, 2021 [1], deal with those models without trying to complete the data but aiming to estimate the relationship between incomplete multidimensional arrays and covariates. This approach is unique and worth research. To handle the unobserved entries, Zhou et al, 2021 [1] consider an optimization problem of a loss function under the low-rankness, sparsity, and fusion constraints.

This approach may have numerous benefits within the context of analyzing biomedical or business data. To enhance the analysis, it is vital to understand, in detail, the algorithms proposed by Zhou et al, 2021 [1] and their theoretical proofs. The nonconvexity of the problem causes the theoretical explanation to be highly nontrivial. To apply those algorithms for a more general model, some of the assumptions about the data could be dropped. Through theoretical analysis, one assumption was weakened. Also, it is necessary to apply those algorithms for a simulation and a real-life dataset. Two data patterns were considered for the simulations. Both illustrated the efficiency and accuracy of the algorithms. Simulations and theoretical proofs show that the estimation error decreases when the observation probability increases. The method is also used to analyze a speech dataset of Parkinson's patients Tsanas et al, 2009 [8]. Several patterns were found by this analysis, which were found consistent existing research on speech analysis for Parkinson's diagnosis.

Overall, the theoretical explanation of proposed in Zhou et al, 2021 [1] algorithms are evaluated, and the results for simulations and a real-life dataset for voice analysis of Parkinson's disease patients are presented.

2 Definitions and notations

In this chapter, we introduce main notations and definitions that are used regularly through the paper. Tensors are multidimensional arrays. For example, matrices are 2-dimensional tensors and MRI images are 3-dimensional tensors. For a m-dimensional tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$: $\mathcal{A}_{i_1,\dots i_m}$ is its $(i_1, \dots i_m)$th entry and $\mathcal{A}_{i_1,\dots i_{j-1},:,i_{j+1},\dots,i_m} = (\mathcal{A}_{i_1,\dots i_{j-1},1,,i_{j+1},\dots,i_m}, \dots, \mathcal{A}_{i_1,\dots i_{j-1},,d_j,i_{j+1},\dots,i_m})^T \in \mathbb{R}^{d_j}$. Let $\text{unfold}_m(\mathcal{A})$ denote the mode-m unfolding of $\mathcal{A}$. Tensor unfolding is also called tensor matricization. For example, the mode-3 unfolding of a third-order tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is $\text{unfold}_3(\mathcal{A}) = [\mathcal{A}_{1,1,:}, \dots, \mathcal{A}_{d_1,1,:}, \dots, \mathcal{A}_{d_1,d_2,:}] \in \mathbb{R}^{(d_1 d_2) \times d_3}$. Let $[d] = \{1, \dots, d\}$. For $a \in \mathbb{R}^{d_j}$, j-mode tensor product is defined as $\mathcal{A} \times_j a \in \mathbb{R}^{d_1 \times \cdots \times d_{j-1} \times d_{j+1} \times \cdots \times d_m}$, such that $(\mathcal{A} \times_j a)_{i_1 \times \cdots \times i_{j-1} \times i_{j+1} \times \cdots \times i_m} = \sum_{i=1}^{d_j} \mathcal{A}_{i_1,\dots,i_m} a_{i_j}$. For $B \in \mathbb{R}^{J \times d_j}$, j-mode matrix product of a tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ is defined as $\mathcal{A} \times_j B \in \mathbb{R}^{d_1 \times \cdots \times \times d_{j-1} \times J \times d_{j+1} \times \cdots \times d_m}$, such that for all $i_1 \in [d_1], \dots i_{j-1} \in [d_{j-1}], i_j \in [J], i_{j+1} \in [d_{j+1}], \dots i_m \in [d_m]$, $(\mathcal{A} \times_j B)_{i_1,\dots i_{j-1},i_j,i_{j+1},\dots i_m} = \sum_{k=1}^{d_j} a_{i_1,\dots,i_m} b_{i_j,k}$. For $a_j \in \mathbb{R}^{d_j}, j \in [m]$, the multilinear combination of the tensor entries is defined as $\mathcal{A} \times_1 a_1 \times_2 \cdots \times_m a_m$ $= \sum_{i_1 \in [d_1]} \cdots \sum_{i_m \in [d_m]} a_{1,i_1} \dots a_{m,i_m} \mathcal{A}_{i_1,\dots,i_m}$. The tensor spectral norm is defined as $\|A\| = \sup_{\|a_1\|=\cdots=\|a_m\|=1} |\mathcal{A} \times_1 a_1 \times_2 \cdots \times_m a_m|$. The tensor Frobenius norm is $\|\mathcal{A}\|_F = \sqrt{\sum_{i_1,\dots,i_m} \mathcal{A}^2_{i_1,\dots,i_m}}$.

Let $\circ$ and $\otimes$ denote outer product and Kronecker product, respectfully. For tensors $\mathcal{A}_1, \dots \mathcal{A}_m$, recall that $\|\mathcal{A}_1 \circ \mathcal{A}_2 \circ \cdots \circ \mathcal{A}_m\|_F^2 = \|\mathcal{A}_1\|_F^2 \|\mathcal{A}_2\|_F^2 \dots \|\mathcal{A}_m\|_F^2$. For a vector $a \in \mathbb{R}^d$, let $\|a\|$ and $\|a\|_0$ denote its Euclidean norm and l_0 norm, respectfully. For vectors a, b, recall the triangle inequality $\|a+b\| \leq \|a\|+\|b\|$ and $\|a-b\| \geq |\|a\|-\|b\||$. For a matrix A, if $Av = \lambda v$, where v is a vector and λ is a scalar. Then v is an eigenvector of A and λ is an eigenvalue corresponding to that eigenvector. Singular value of A is defined as the square root of the non-negative eigenvalue of the matrix A^*A, where A^* denotes the conjugate transpose of A. Since $A \in \mathbb{R}^d$, $A^* = A^T$.

For a matrix $A \in \mathbb{R}^{d_1 \times d_2}$, let $\|A\|$ denote its spectral norm as

$$\|A\| = \sqrt{\lambda_{\max}(A^*A)} = \sigma_{\max}(A),$$

where $\sigma_{\max}(A)$ is the largest singular value of matrix A. For A, B - matrices, recall that $\|AB\| \leq \|A\|\|B\|$. Let $\psi_2 = \exp(x^2) - 1$. Then, for a random variable X, let $\|X\|_{\psi_2}$ denote its Orlicz norm, defined as $\inf \left\{u > 0 : \mathbb{E}\left[\psi_2\left(\frac{|X|}{u}\right)\right] \leq 1\right\} = \left\{u > 0 : \mathbb{E}\left[\exp\left(\frac{|X|^2}{u^2}\right)\right] \leq 2\right\}$.

3 Statistical model

In this section, we introduce the tensor regression model that is a foundation for the algorithms of our interest and their analysis. Also, we present the assumptions on the coefficient tensor and establish an optimization problem.

At each time point t an mth-order tensor Y_t of dimension $d_1 \times \cdots \times d_m, t \in [T]$ is collected. Suppose there are n subjects in the study. For each subject i, this tensor sequence can be represented as a **dynamic** $(m+1)$th-order tensor Y_i of dimension $d_1 \times \cdots \times d_m \times T$.

A **dynamic** tensor Y_i and a q-dimensional vector of covariates x_i are collected, where

$$Y_i \in \mathbb{R}^{d_1 \times \cdots \times d_m \times T} \text{ and } x_i \in \mathbb{R}^q, i \in [n].$$

The response tensor Y_i can be partially observed, with a missing pattern varying from subject to subject. We consider the following tensor regression model:

$$Y_i = B^* \times_{m+2} x_i + \mathcal{E}_i \tag{1}$$

$B^* \in \mathbb{R}^{d_1 \times \cdots \times d_m \times T \times q} - (m+2)$th-order coefficient tensor,
$\mathcal{E}_i \in \mathbb{R}^{d_1 \times \cdots \times d_m \times T} - (m+1)$th-order error tensor independent of x_i. Without loss of generality, we assume that the response tensor is centered. Therefore, the intercept from the model can be dropped.

The main object of interest in the analysis is to estimate the coefficient tensor B^*.

Assumptions on B^*:

1. B^* admits a rank-r CP decomposition structure.

$$B^* = \sum_{k \in [r]} w_k^* \beta_{k,1}^* \circ \cdots \circ \beta_{k,m+2}^* \text{ , where} \tag{2}$$

   $w_k^* > 0$ and $\beta_{k,j}^* \in \mathbb{S}^{d_j}, \mathbb{S}^{d_j} = \{a \in \mathbb{R}^d \mid \|a\| = 1\}$.

2. B^* is sparse, thus the decomposed components $\beta_{k,j}^*$'s are sparse too.
   $\beta_{k,j}^* \in S(d_j, s_j)$ for all $j \in [m+1], k \in [r]$, where

$$S(d,s) = \left\{ \beta \in \mathbb{R} \mid \sum_{l=1}^{d} \mathbb{1}_{(\beta_l \neq 0)} \leq s \right\} = \left\{ \beta \in \mathbb{R} \mid \|\beta\|_0 \leq s \right\}.$$

   This assumption enables to concentrate on the tensor regions that are the most dependent on the covariates.

3. Decomposed components $\beta_{k,j}^*$ have fusion structure,
   $\beta_{k,j}^* \in F(d_j, f_j)$ for all $j \in [m+1], k \in [r]$, where

$$F(d,f) = \left\{ \beta \in \mathbb{R} \mid \sum_{l=2}^{d} \mathbb{1}_{|\beta_l - \beta_{l-1}| \neq 0)} \leq f \right\} = \left\{ \beta \in \mathbb{R} \mid \|D\beta\|_0 \leq f - 1 \right\} \text{ , where}$$

$D \in \mathbb{R}^{(d-1)\times d}$ with $D_{i,i} = -1, D_{i,i+1} = 1$ for $i \in [d-1]$ and other entries $= 0$.

This assumption encourages temporal smoothness and helps pool information from tensors observed at adjacent time points.

A major challenge in this model is that some entries of the tensor Y are unobserved. Let $\Omega_i \subseteq [d_1] \times [d_2] \times \cdots \times [d_{m+1}]$ denote the set of indexes for the observed entries in $Y_i, i \in [n]$. Also, a projection tensor $\Pi_\Omega(\cdot)$ is defined as:

$$[\Pi_\Omega(Y)]_{i_1,i_2,\ldots,i_{m+1}} = \begin{cases} Y_{i_1,i_2,\ldots,i_{m+1}} & \text{if}(i_1, i_2, \ldots, i_{m+1}) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

Note that for tensors $\mathcal{A}_1, \ldots \mathcal{A}_m$,

$$\|\Pi_\Omega(\mathcal{A}_1 \circ \mathcal{A}_2 \circ \cdots \circ \mathcal{A}_m)\|_F^2 = \|\Pi_\Omega(\mathcal{A}_1)\|_F^2 \|\mathcal{A}_2\|_F^2 \ldots \|\mathcal{A}_m\|_F^2$$

$$= \|\mathcal{A}_1\|_F^2 \|\Pi_\Omega(\mathcal{A}_2)\|_F^2 \ldots \|\mathcal{A}_m\|_F^2 = \|\mathcal{A}_1\|_F^2 \|\mathcal{A}_2\|_F^2 \ldots \|\Pi_\Omega(\mathcal{A}_m)\|_F^2.$$

Also, for tensors $\mathcal{A}_1, \mathcal{A}_2$, we get

$$\langle \Pi_\Omega(\mathcal{A}_1), \Pi_\Omega(\mathcal{A}_2) \rangle = \langle \Pi_\Omega(\mathcal{A}_1), \mathcal{A}_2 \rangle.$$

Then, we consider the following constrained optimization problem:

$$\min_{w_k, \beta_{k,j} k\in[r], j\in[m+2]} \frac{1}{n} \sum_{i=1}^{n} \left\| \Pi_{\Omega_i} \left(Y_i - \sum_{k\in[r]} w_k (\beta_{k,m+2}^T x_i) \beta_{k,1} \circ \cdots \circ \beta_{k,m+1} \right) \right\|_F^2 \tag{3}$$

with limitations $\|\beta_{k,j}\|_2 = 1, j \in [m+2], \|\beta_{k,j}\|_0 \le \tau_{s_j}, \|D\beta_{k,j}\|_0 \le \tau_{f_j}$, $j \in [m+1], k \in [r]$.

4 Estimation methods and algorithms

In this chapter, we present algorithms for estimating a solution to the optimization problem (3) and initializing the variables to achieve more precise results.

4.1 Estimation algorithm

In this section, we introduce an algorithm to solve the optimization problem (3). We go step by step to understand in detail and derive the formulas used in the algorithm.

The optimization problem (3) is a non-convex optimization with multiple constraints. The loss function is non-trivial since a projection tensor was added to deal with unobserved entries. Problem (3) either does not have a closed-form solution or it is too complex for calculations. Therefore, estimation algorithms should be considered to find a solution of (3).

Zhou et al, 2021 [1] proposed an alternating block updating algorithm to solve this optimization problem.

Algorithm 1 Alternating block updating algorithm

Input: the data $\{(x_i, Y_i, \Omega_i), i = 1, \ldots, n\}$, the rank r, the sparsity parameter τ_{s_j}, and the fusion parameter $\tau_{f_j}, j \in [m+1]$.
Initialization: set $w_k = 1$, and randomly generate unit norm vectors $\beta_{k,1}, \ldots, \beta_{k,m+2}$ from a standard normal distribution, $k \in [r]$.
Repeat
for $k = 1$ to r **do**
  for $j = 1$ to $m+1$ **do**
    Step 1: obtain the unconstrained estimator $\tilde{\beta}_{k,j}^{(t+1)}$, given, $\hat{w}_k^{(t)}, \hat{\beta}_{k,1}^{(t+1)}, \ldots, \hat{\beta}_{k,j-1}^{(t+1)}, \hat{\beta}_{k,j+1}^{(t)}, \ldots, \hat{\beta}_{k,m+1}^{(t)}, \hat{\beta}_{k,m+2}^{(t)}$, by solving (4).
    Normalize $\tilde{\beta}_{k,j}^{(t+1)}$.
    Step 2: obtain the constrained estimator $\hat{\beta}_{k,j}^{(t+1)}$, by applying the *Truncatefuse* operator to $\tilde{\beta}_{k,j}^{(t+1)}$
    Normalize $\hat{\beta}_{k,j}^{(t+1)}$.
  end for
  Step 3: obtain $\tilde{w}_k^{(t+1)}$, given $\hat{\beta}_{k,1}^{(t+1)}, \ldots, \hat{\beta}_{k,m+1}^{(t+1)}, \hat{\beta}_{k,m+2}^{(t)}$ using (8)
  Step 4: obtain $\tilde{\beta}_{k,m+2}^{(t+1)}$, given $\hat{w}_k^{(t+1)}, \hat{\beta}_{k,1}^{(t+1)}, \ldots, \hat{\beta}_{k,m+1}^{(t+1)}$ using (9)
end for
Until the stopping criteria is met.
Output: $\hat{w}_k, \hat{\beta}_{k,1}, \ldots, \hat{\beta}_{k,m+2}, k \in [r]$.

Step 1: solving an unconstrained weighted tensor completion problem

$$\min_{\beta_{k,j}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \alpha_{i,k}^{(t)} \right\} \left\| \Pi_{\Omega_i} \left(R_{i,k}^{(t+1)} - \hat{w}_k \hat{\beta}_{k,1}^{(t+1)} \circ \cdots \circ \hat{\beta}_{k,j-1}^{(t+1)} \circ \beta_{k,j} \circ \hat{\beta}_{k,j+1}^{(t)} \circ \cdots \circ \hat{\beta}_{k,m+1}^{(t)} \right) \right\|_F^2, \tag{4}$$

where $\alpha_{i,k}^{(t)} = \beta_{k,m+2}^{(t)T} x_i$ and $R_{i,k}^{(t+1)}$ is a residual form defined as

$$R_{i,k}^{(t+1)} = \frac{\left(Y_i - \sum\limits_{k'<k} \hat{w}_{k'}^{(t+1)} \alpha_{i,k'}^{(t+1)} \beta_{k',1}^{(t+1)} \circ \cdots \circ \beta_{k',m+1}^{(t+1)} - \sum\limits_{k'>k} \hat{w}_{k'}^{(t)} \alpha_{i,k'}^{(t)} \beta_{k',1}^{(t)} \circ \cdots \circ \beta_{k',m+1}^{(t)}\right)}{\alpha_{i,k}^{(t)}} \tag{5}$$

$$\text{for } i \in [n], k \in [r].$$

The optimization problem (4) has a closed-form solution.To simplify the calculations, the solution is presented for $m = 2$. For $m \geq 3$ calculations are similar. In particular, $\tilde{\beta}_{k,3}$ is estimated as follows:

The optimization problem becomes:

$$\min_{\beta_{k,3}} \frac{1}{n} \sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\} \left\| \Pi_{\Omega_i} \left(R_{i,k}^{(t+1)} - \hat{w}_k \hat{\beta}_{k,1}^{(t+1)} \circ \hat{\beta}_{k,2}^{(t+1)} \circ \beta_{k,3} \right) \right\|_F^2$$

$$\left\| \Pi_{\Omega_i} \left(R_{i,k}^{(t+1)} - \hat{w}_k \hat{\beta}_{k,1}^{(t+1)} \circ \hat{\beta}_{k,2}^{(t+1)} \circ \beta_{k,3} \right) \right\|_F^2$$

$$= \sum_{l_1,l_2,l} \left| \delta_{i,l_1,l_2,l} \left(R_{i,k,l_1,l_2,l}^{(t+1)} - \hat{w}_k \hat{\beta}_{k,1,l_1}^{(t+1)} \hat{\beta}_{k,2,l_2}^{(t+1)} \beta_{k,3,l} \right) \right|^2,$$

where $\delta_{i,l_1,l_2,l}$ is an indicator function on Ω_i, which means that $\delta_{i,l_1,l_2,l} = 1$ if $(l_1, l_2, l) \in \Omega_i$ and $\delta_{i,l_1,l_2,l} = 0$ otherwise.

The function can be minimized by every entry of $\beta_{k,3}$:

$$\sum_{i=1}^{n} (\alpha_{i,k}^{(t)})^2 \sum_{l_1,l_2} \left| \delta_{i,l_1,l_2,l} \left(R_{i,k,l_1,l_2,l}^{(t+1)} - \hat{w}_k \hat{\beta}_{k,1,l_1}^{(t+1)} \hat{\beta}_{k,2,l_2}^{(t+1)} \beta_{k,3,l} \right) \right|^2 \to \min$$

Let $\alpha_i = (\alpha_{i,k}^{(t)})^2, (l_1, l_2) = j, \beta_{k,3,l} = x, \delta_{i,l_1,l_2,l} R_{i,k,l_1,l_2,l}^{(t+1)} = a_{i,j}$
and $\delta_{i,l_1,l_2,l} \hat{w}_k \hat{\beta}_{k,1,l_1}^{(t+1)} \hat{\beta}_{k,2,l_2}^{(t+1)} = b_{i,j}$.
Hence, by applying Proposition A.1, we get:

$$\tilde{\beta}_{k,3,l} = \frac{\sum\limits_{i=1}^{n} (\alpha_{i,k}^{(t)})^2 \sum\limits_{l_1,l_2} \delta_{i,l_1,l_2,l} R_{i,k,l_1,l_2,l}^{(t+1)} \hat{\beta}_{k,1,l_1}^{(t+1)} \hat{\beta}_{k,2,l_2}^{(t+1)}}{\sum\limits_{i=1}^{n} (\alpha_{i,k}^{(t)})^2 \sum\limits_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{w}_k (\hat{\beta}_{k,1,l_1}^{(t+1)})^2 (\hat{\beta}_{k,2,l_2}^{(t+1)})^2}. \tag{6}$$

Similarly, we get a closed form solutions for $\tilde{\beta}_{k,2,l}$ and $\tilde{\beta}_{k,1,l}$

$$\tilde{\beta}_{k,2,l} = \frac{\sum\limits_{i=1}^{n} (\alpha_{i,k}^{(t)})^2 \sum\limits_{l_1,l_3} \delta_{i,l_1,l_3,l} R_{i,k,l_1,l_3,l}^{(t+1)} \hat{\beta}_{k,1,l_1}^{(t+1)} \hat{\beta}_{k,3,l_3}^{(t+1)}}{\sum\limits_{i=1}^{n} (\alpha_{i,k}^{(t)})^2 \sum\limits_{l_1,l_3} \delta_{i,l_1,l_3,l} \hat{w}_k (\hat{\beta}_{k,1,l_1}^{(t+1)})^2 (\hat{\beta}_{k,3,l_3}^{(t+1)})^2}.$$

$$\tilde{\beta}_{k,1,l} = \frac{\sum_{i=1}^{n}(\alpha_{i,k}^{(t)})^2 \sum_{l_2,l_3} \delta_{i,l_2,l_3,l} R_{i,k,l_2,l_3,l}^{(t+1)} \hat{\beta}_{k,2,l_2}^{(t+1)} \hat{\beta}_{k,3,l_3}^{(t+1)}}{\sum_{i=1}^{n}(\alpha_{i,k}^{(t)})^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{w}_k (\hat{\beta}_{k,1,l_1}^{(t+1)})^2 (\hat{\beta}_{k,2,l_2}^{(t+1)})^2}.$$

As seen from the formulas above, $\tilde{\beta}$'s need to be updated in an element-wise fashion as an indicator δ_{i,l_1,l_2,l_3} is present in both nominator and denominator. Since it could change across different entries of $\tilde{\beta}$'s, it can not be canceled.

Step 2: applying operators to the unconstrained estimators to incorporate the sparsity and fusion constraints

As in Zhou et al, 2021 [1], we define *Truncate* operator as follows

$$[Truncate(a, \tau_s)]_j = \begin{cases} a_j & \text{if } j \in \text{supp}(a, \tau_s) \\ 0 & \text{otherwise,} \end{cases}$$

where $\text{supp}(a, \tau_s)$ refers to the indexes of τ_s entries with the largest absolute values in a. The truncation operator ensures that the total number of nonzero entries in a is bounded by τ_s.

We also consider *Fuse* operator which is defined as

$$[Fuse(a, \tau_f)]_j = \sum_{i=1}^{\tau_f} \mathbb{1}_{j \in C_i} \frac{1}{|C_i|} \sum_{l \in C_i} a_l,$$

where $\{C_i\}_{i=1}^{\tau_f}$ are the fusion groups. To calculate the fusion groups:

1. Calculate $Truncate(Da, \tau_f - 1)$. The resulting vector has at most $\tau_f - 1$ nonzero entries.
2. The elements a_j and a_{j+1} are put into the same group if $[Truncate(Da, \tau_f - 1)]_j = 0$.
3. Elements of each group are averaged.

Combining *Truncate* and *Fuse* operators, we obtain *Truncasefuse* operator, defined as

$$Truncatefuse(a, \tau_s, \tau_f) = Truncate\left[Fuse(a, \tau_f), \tau_s\right].$$

For example, consider $a = (0.2, 0.1, 0.5, 0.6, 0.7)^T, \tau_s = 3$ and $\tau_f = 2$.

1. $Da = (-0.1, 0.4, 0.1, 0.1)^T$.
2. $Truncate(Da, \tau_f - 1) = Truncate((-0.1, 0.4, 0.1, 0.1)^T, 1) = (0, 0.4, 0.1, 0.1)^T$.
3. The previous step shows that a_1, a_2 belong to one group, and a_3, a_4, a_5 belong to the other.

4. *Fuse* $(a, \tau_f) = (0.15, 0.15, 0.6, 0.6, 0.6)^T$.

5. *Truncatefuse*(a, τ_s, τ_f) = *Truncate* $[$*Fuse*$(a, 2), 3]$ = *Truncate* $\left[(0, 0, 0.6, 0.6, 0.6)^T\right]$.

Step 3: update $\hat{w}_k^{(t+1)}$
Since $\hat{\beta}_{k,1}^{(t+1)}, \ldots, \hat{\beta}_{k,m+1}^{(t+1)}$ are already estimated, the optimization problem becomes:

$$\min_{w_k} \frac{1}{n} \sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \left\| \Pi_{\Omega_i} \left(R_{i,k}^{(t+1)} - w_k \hat{\beta}_{k,1}^{(t+1)} \circ \cdots \circ \hat{\beta}_{k,m+1}^{(t+1)} \right) \right\|_F^2,$$

where

$$R_{i,k}^{(t+1)} = \frac{\left(Y_i - \sum_{k'<k} w_{k'}^{(t+1)} \alpha_{i,k'}^{(t)} \beta_{k',1}^{(t+1)} \circ \cdots \circ \beta_{k',m+1}^{(t+1)} - \sum_{k'>k} w_{k'}^{(t)} \alpha_{i,k'}^{(t)} \beta_{k',1}^{(t+1)} \circ \cdots \circ \beta_{k',m+1}^{(t+1)} \right)}{\alpha_{i,k}^{(t)}}.$$

Note that $R_{i,k}^{(t+1)}$ does not depend on w_k. By the definition of the Frobenius norm, we have

$$\left\| \Pi_{\Omega_i} \left(R_{i,k}^{(t+1)} - w_k \hat{\beta}_{k,1}^{(t+1)} \cdots \hat{\beta}_{k,m+1}^{(t+1)} \right) \right\|_F^2$$

$$= \sum_{l_1, \ldots, l_{m+1}} \left| \delta_{i,l_1,\ldots,l_{m+1}} \left(R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)} - w_k \hat{\beta}_{k,1,l_1}^{(t+1)} \cdots \hat{\beta}_{k,m+1,l_{m+1}}^{(t+1)} \right) \right|^2.$$

Then, the function we minimize in w_k becomes:

$$\frac{1}{n} \sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \sum_{l_1, \ldots, l_{m+1}} \left| \delta_{i,l_1,\ldots,l_{m+1}} \left(R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)} - w_k \hat{\beta}_{k,1,l_1}^{(t+1)} \cdots \hat{\beta}_{k,m+1,l_{m+1}}^{(t+1)} \right) \right|^2$$

Let $x = w_k$, $\alpha_i = \left\{\alpha_{i,k}^{(t)}\right\}^2$, $j = (l_1, \ldots, l_{m+1})$, $a_{i,j} = \delta_{i,l_1,\ldots,l_{m+1}} R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)}$ and $b_{i,j} = \delta_{i,l_1,\ldots,l_{m+1}} \hat{\beta}_{k,1,l_1}^{(t+1)} \cdots \hat{\beta}_{k,m+1,l_{m+1}}^{(t+1)}$. Therefore, by applying Proposition A.1, the closed form solution is:

$$\hat{w}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \sum_{l_1, \ldots, l_{m+1}} \delta_{i,l_1,\ldots,l_{m+1}} R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)} \hat{\beta}_{k,1,l_1}^{(t+1)} \cdots \hat{\beta}_{k,m+1,l_{m+1}}^{(t+1)}}{\sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \sum_{l_1, \ldots, l_{m+1}} \left(\delta_{i,l_1,\ldots,l_{m+1}} \hat{\beta}_{k,1,l_1}^{(t+1)} \cdots \hat{\beta}_{k,m+1,l_{m+1}}^{(t+1)} \right)^2}.$$

Then,

$$\hat{w}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \Pi_{\Omega_i}(R_{i,k}^{(t+1)}) \times_1 \hat{\beta}_{k,1}^{(t+1)} \times_2 \cdots \times_{m+1} \hat{\beta}_{k,m+1}^{(t+1)}}{\sum_{i=1}^{n} \left\{\alpha_{i,k}^{(t)}\right\}^2 \left\| \Pi_{\Omega_i} \left(\hat{\beta}_{k,1}^{(t+1)} \circ \cdots \circ \hat{\beta}_{k,m+1}^{(t+1)} \right) \right\|_F^2}. \tag{7}$$

Step 4: update $\hat{\beta}_{k,m+2}^{(t+1)}$

Since $\hat{\beta}_{k,1}^{(t+1)}, \ldots, \hat{\beta}_{k,m+1}^{(t+1)}$ and $\hat{w}_k^{(t+1)}$ are already estimated, the optimization problem becomes:

$$\min_{\beta_{k,m+2}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \beta_{k,m+2}^T x_i \right\}^2 \left\| \Pi_{\Omega_i} \left(\frac{R_{i,k}^{(t+1)}}{\beta_{k,m+2}^T x_i} - A_k^{(t+1)} \right) \right\|_F^2,$$

where

$$R_{i,k}^{(t+1)} = Y_i - \sum_{k' \neq k, k' \in [r]} \hat{w}_{k'}^{(t+1)} \beta_{k',m+2}^{(t)T} x_i \beta_{k',1}^{(t+1)} \circ \cdots \circ \beta_{k',m+1}^{(t+1)} \text{ and}$$

$$A_k^{(t+1)} = \hat{w}_k^{(t+1)} \hat{\beta}_{k,1}^{(t+1)} \circ \cdots \circ \hat{\beta}_{k,m+1}^{(t+1)}.$$

By the definition of the Frobenius norm, we have

$$\left\| \Pi_{\Omega_i} \left(\frac{R_{i,k}^{(t+1)}}{\beta_{k,m+2}^T x_i} - A_k^{(t+1)} \right) \right\|_F^2 = \sum_{l_1, \ldots, l_{m+1}} \left| \delta_{i,l_1,\ldots,l_{m+1}} \left(\frac{R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)}}{\beta_{k,m+2}^T x_i} - A_{k,l_1,\ldots,l_{m+1}}^{(t+1)} \right) \right|^2.$$

Therefore, the function that needs to be minimized in $\beta_{k,m+2}$ becomes

$$\sum_{i=1}^{n} \sum_{l_1, \ldots, l_{m+1}} \left(\delta_{i,l_1,\ldots,l_{m+1}} \left(R_{i,k,l_1,\ldots,l_{m+1}}^{(t+1)} - A_{k,l_1,\ldots,l_{m+1}}^{(t+1)} \beta_{k,m+2}^T x_i \right) \right)^2.$$

Using Proposition A.1, a closed form solution is:

$$\hat{\beta}_{k,m+2}^{(t+1)} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\| \Pi_{\Omega_i} \left(A_k^{(t+1)} \right) \right\|_F^2 x_i x_i^T \right\}^{-1} \sum_{i=1}^{n} \left\langle \Pi_{\Omega_i}(R_{i,k}^{(t+1)}), \Pi_{\Omega_i}(A_k^{(t+1)}) \right\rangle x_i. \quad (8)$$

With a good initialization, which can be achieved through the algorithm below, the iterative estimator from the considered algorithm is within the statistical precision of the true parameter. Results from the theorems in the later chapters provide a theoretical condition to end the iterative process: the computation error is dominated by statistical error. In practice, the iteration ends when two consecutive iterations are close.

4.2 Initialization algorithms

The success of the alternating block updating algorithm depends on a good initialization of the main variables. In the section, we consider two initialization algorithms for the cases of $r = 1$ and $r > 1$. Since the optimization problem (3) is non-convex, the initialization might not have a closed-form solution. Therefore, initialization algorithms that solve this issue are considered. Zhou et al, 2021 [1] provide an algorithm for initialization called a spectral initialization. In order to simplify the presentation, the notion is used for the case of $m = 2$. Nevertheless, it can be extended to cases where $m > 2$. Let

$$\mathcal{T} = \frac{1}{n} \sum_{i} \Pi_{\Omega_i}(Y_i),$$

$$A_1 = \text{unfold}_3(p^{-1}\mathcal{T}) \in \mathbb{R}^{d_3 \times d_1 d_2}, A_2 = \text{unfold}_1(p^{-1}\mathcal{T}) \in \mathbb{R}^{d_1 \times d_2 d_3}$$

$$B_1 = \Pi_{\text{off-diag}}(A_1 A_1^T) \in \mathbb{R}^{d_3 \times d_3}, B_2 = \Pi_{\text{off-diag}}(A_2 A_2^T) \in \mathbb{R}^{d_1 \times d_1},$$

where $\Pi_{\text{off-diag}}$ keeps only the off-diagonal entries of the matrix.

Let $U_1 \Lambda_1 U_1^T$ be the rank-r decomposition of B_1, and let $U_2 \Lambda_2 U_2^T$ be the rank-r decomposition of B_2.

Algorithm 2 Spectral initialization algorithm for $r = 1$

Input: the number of restarts L, the estimates U_1, U_2,and the sparsity parameter $\tau_{s_j}, j \in [3]$
for $l = 1$ to L **do**
Generate $g_1^l \sim \text{Normal}(0, I_{d_3})$, and compute $\tilde{g}_1^l = U_1 U_1^T g_1, M_1^l = p^{-1}\mathcal{T} \times_3 \tilde{g}_1^l$.
Set v_1^l and v_2^l are the first left and right singular vector of M_1^l corresponding to the largest absolute value $|\lambda_1^l|$.
end for
for $l = 1$ to L **do**
Generate $g_2^l \sim \text{Normal}(0, I_{d_1})$, and compute $\tilde{g}_2^l = U_2 U_2^T g_2, M_2^l = p^{-1}\mathcal{T} \times_3 \tilde{g}_2^l$.
Set v_3^l and v_4^l are the first left and right singular vector of M_2^l corresponding to the largest absolute value $|\lambda_2^l|$.
end for
Choose (v_1, v_2) from $\{(v_1, v_2)\}_{l=1}^L$ with the largest $|\lambda_1^l|$.
Choose (v_3, v_4) from $\{(v_3, v_4)\}_{l=1}^L$ with the largest $|\lambda_2^l|$.
Compute $\hat{\beta}_{1,j}^{(0)} = \texttt{Norm}(\textit{Truncate}(\tilde{v}_j, \tau_{s_j}))$ for $j = 1, 2, 3$, where $(\tilde{v}_1, \tilde{v}_2, \tilde{v}_3)$ is obtained from $(v_1, v_2), (v_3, v_4)$ and Norm is the normalization operator.
Compute $\hat{w}_1^{(0)}$ and $\hat{\beta}_{1,4}^{(0)}$ using (9).
Output: $\hat{w}_1^{(0)}, \hat{\beta}_{1,1}^{(0)}, \hat{\beta}_{1,2}^{(0)}, \hat{\beta}_{1,3}^{(0)}, \hat{\beta}_{1,4}^{(0)}$.

Given $\hat{\beta}_{1,1}^{(0)}, \hat{\beta}_{1,2}^{(0)}, \hat{\beta}_{1,3}^{(0)}$, we have the following optimization problem,

$$\min_{w_1 > 0, \|\beta_{1,4}\|=1} \frac{1}{n} \sum_{i=1}^n \left\| \Pi_{\Omega_i} \left(Y_i - \hat{w}_k (\beta_{1,4}^T x_i) \hat{\beta}_{1,1}^{(0)} \circ \hat{\beta}_{1,2}^{(0)} \circ \hat{\beta}_{1,3}^{(0)} \right) \right\|_F^2$$

The solutions are obtained as:

$$\hat{\beta}_{1,4}^{(0)} = \left\{ \frac{1}{n} \sum_{i=1}^n \left\| \Pi_{\Omega_i} \left(\hat{\beta}_{1,1}^{(0)} \circ \hat{\beta}_{1,2}^{(0)} \circ \hat{\beta}_{1,3}^{(0)} \right) \right\|_F^2 x_i x_i^T \right\}^{-1} n^{-1}$$

$$\times \sum_{i=1}^n \left\langle \Pi_{\Omega_i}(Y_i), \Pi_{\Omega_i}(\hat{\beta}_{1,1}^{(0)} \circ \hat{\beta}_{1,2}^{(0)} \circ \hat{\beta}_{1,3}^{(0)}) \right\rangle x_i,$$

$$\hat{w}_1^{(0)} = \frac{\sum_{i=1}^n \hat{\beta}_{1,4}^T x_i \Pi_{\Omega_i}(Y_i) \times_1 \hat{\beta}_{1,1}^{(0)} \times_2 \hat{\beta}_{1,2}^{(0)} \times_3 \hat{\beta}_{1,3}^{(0)}}{\sum_{i=1}^n \left\{ \hat{\beta}_{1,4}^T x_i \right\}^2 \left\| \Pi_{\Omega_i} \left(\hat{\beta}_{1,1}^{(0)} \circ \hat{\beta}_{1,2}^{(0)} \circ \hat{\beta}_{1,3}^{(0)} \right) \right\|_F^2} \tag{9}$$

Zhou et al, 2021 [1] also propose the initialization algorithm for a more general case where $r > 1$:

Algorithm 3 Spectral initialization algorithm for $r > 1$

Input: the number of restarts L, the estimates U_1, U_2, the tolerance parameter ϵ_{th} and the sparsity parameter $\tau_{s_j}, j \in [3]$

Obtain $\{(v_1, v_2)\}_{l=1}^{L}$ and $\{(v_3, v_4)\}_{l=1}^{L}$ using Algorithm 2.

Obtain the triplet $S = \{(\tilde{v}_1, \tilde{v}_2, \tilde{v}_3)\}_{l=1}^{L}$ from $(v_1, v_2)_{l=1}^{L}, (v_3, v_4)_{l=1}^{L}$

for $l = 1$ to L **do**

Find $(\hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \hat{\beta}_{k,3}) = \text{argmax}_{(\tilde{v}_1, \tilde{v}_2, \tilde{v}_3)) \in S} |p^{-1}\mathcal{T} \times_1 \tilde{v}_1 \times_2 \tilde{v}_2 \times_3 \tilde{v}_3|$

Remove all the triplets in $(\tilde{v}_1, \tilde{v}_2, \tilde{v}_3))_{l=1}^{L}$ with $\max\{|\langle \hat{\beta}_{k,1}, \tilde{v}_1^l \rangle|, |\langle \hat{\beta}_{k,2}, \tilde{v}_2^l \rangle|, |\langle \hat{\beta}_{k,3}, \tilde{v}_3^l \rangle|\} > 1 - \epsilon_{\text{th}}$

end for

Set $\hat{w}_k = 1$, and randomly generate unit-norm vectors $\hat{\beta}_{k,4}, k \in [r]$ from a standard normal distribution.

Repeat

Update $\hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \hat{\beta}_{k,3}$ using (6), and set $\hat{\beta}_{k,j} = \texttt{Norm}(\texttt{Truncate}(\hat{\beta}_{k,j}, \tau_{s_j}))$ for $j = 1, 2, 3$.

Update $\hat{w}_k$ using (7).

Update $\hat{\beta}_{k,4}$ using (8).

Until the stopping criteria is met.

Denote the final update of $\hat{w}_k, \hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \hat{\beta}_{k,3}, \hat{\beta}_{k,4}$ as $\hat{w}_k^{(0)}, \hat{\beta}_{k,1}^{(0)}, \hat{\beta}_{k,2}^{(0)}, \hat{\beta}_{k,3}^{(0)}, \hat{\beta}_{k,4}^{(0)}, k \in [r]$, respectively.

Output: $\hat{w}_k^{(0)}, \hat{\beta}_{k,1}^{(0)}, \hat{\beta}_{k,2}^{(0)}, \hat{\beta}_{k,3}^{(0)}, \hat{\beta}_{k,4}^{(0)}, k \in [r]$.

5 Theoretical results

In this chapter, we discuss assumptions that are implemented for a theoretical analysis of Algorithms 1 and 2. We present two theorems that show that estimation from Algorithm 1 gives precise results with a high probability. Also, we present a theorem to theoretically prove that Algorithms 2 provides a good initialization.

5.1 Assumptions

In this section, we introduce assumptions on the statistical model (1). Precise results of Algorithms 1 and 2 cannot be achieved using any initial parameters or under any true model. Therefore, several limitations and assumptions are implemented. We discuss those assumptions and point at the ones that could be weakened. At first, the general assumptions are discussed.

Assumption 1.

(a) The predictor x_i satisfies:

$$\|x_i\| \leq c_1 \text{ and } \frac{1}{n}\sum_{i=1}^{n} \|x_i x_i^T\|_2 \leq c_2, i \in [n]$$

$1/c_0 < \lambda_{\min} \leq \lambda_{\max} < c_0$, where $\lambda_{\min}, \lambda_{\max}$ are the minimum and maximum eigenvalues of the sample covariance matrix $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$, respectively, and c_0, c_1, c_2 are some positive constants.

(b) The true coefficient tensor B^* satisfies the CP decomposition with sparsity and fusion constraints. The decomposition is unique up to a permutation.

(c) The decomposed component $\beta^*_{k,j}$ is a μ-mass unit vector:

$$\max_{l \in d_j} |\beta^*_{k,j,l}| \leq \mu/\sqrt{s} \leq 1.$$

(d) The entries in the error tensor $\mathcal{E}_i$ are independent and identically distributed sub-Gaussian with a variance σ^2.

(e) The entries of the response tensor Y_i are observed independently with an equal probability $p \in (0, 1]$.

Remarks:

1. In Zhou et al, 2021 [1], the assumption in (b) is added with $\|B^*\| \leq c_3 w^*_{\max}$ where $w^*_{\max} = \max_k\{w^*_k\}, w^*_{\min} = \min_k\{w^*_k\}$. Also, $w^*_{\max} = \mathcal{O}(w^*_{\min})$. However, this assumption is not used in the proofs.

2. In real applications, assumption (e) might not be satisfied due to the nature of the unobserved entries. This assumption is required for the theoretical analysis, even though the algorithm does not require it.

Then, we introduce regularity conditions for $r = 1$.
Assumption 2.
Let $d = \max\{d_1, \cdots d_{m+1}\}$

(a) The observation probability p satisfies:

$$p \geq \frac{c_4\{\log(d)\}^4\mu^3}{ns^{1.5}},$$

where c_4 is some positive constant.

(b) The sparsity and fusion parameters satisfy:

$$\tau_{s_j} \geq s_j, \tau_{s_j} = \int_{|} \text{ and } \tau_{f_j} \geq f_j$$

Also, for the minimal gap $\Delta^* = \min_{1<s\leq d_j, \beta^*_{1,j,s} \neq \beta^*_{1,j,s-1}, j\in[3]} |\beta^*_{1,j,s} - \beta^*_{1,j,s-1}|$:

$$\Delta^* > \frac{C_1\sigma}{w_1^*}\sqrt{\frac{s\log(d)}{np}},$$

where C_1 is the positive constant defined in Theorem 5.1.

(c) For the initialization error $\epsilon = \max\{|\hat{w}_1^{(0)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(0)} - \beta_{1,j}^*\|_2\}$:

$$\epsilon < \min\left\{\frac{\lambda_{\min}^3}{24\sqrt{10}c_2\lambda_{\max}^2}, \frac{1}{6}\right\},$$

where c_2 is the same constant as in Assumption 1.

(d) The sample size n satisfies:

$$n \geq \min\left\{\frac{c_5\sigma^2 s^2\log(d)}{w_1^{*2}p}, \frac{c_6\sigma s\log(\sqrt{s^3/p})}{w_1^* p}\right\},$$

where c_5 and c_6 are some positive constants.

Remarks:

1. The condition (a) places a lower bound on the observation probability to ensure a recovery of the tensor coefficient.

2. The condition (b) for the sparsity parameter ensures that the truly nonzero elements would not shrunk to zero.

3. The assumption (c) requires the initial values to be reasonable close to the true parameters, which can be achieved with the considered above initialization algorithm. The condition on the minimal gap ensure that the fused estimator wouldn't incorrectly merge two distinct groups.

4. For the most part, those assumption have a theoretical explanation. Indeed, they are used in the proof of the Theorem 5.1.

Finally, we introduce regularity conditions for $r > 1$.
Assumption 3.

(a) The observation probability p satisfies:

$$p \geq \frac{c_7\{\log(d)\}^4\mu^3 r w_{\max}^{*2}}{ns^{1.5}w_{\min}^{*2}},$$

where c_7 is some positive constant.

(b) The sparsity and fusion parameters satisfy:

$$\tau_{s_j} \geq s_j, \tau_{s_j} = \int_{|} \text{ and } \tau_{f_j} \geq f_j$$

Also, for the minimal gap $\Delta^* = \min_{1<s\leq d_j, \beta^*_{1,j,s}\neq\beta^*_{1,j,s-1}, j\in[3]} |\beta^*_{1,j,s} - \beta^*_{1,j,s-1}|$:

$$\Delta^* > \frac{C_1\sigma}{w_1^*}\sqrt{\frac{s\log(d)}{np}},$$

where C_1 is the positive constant.

(c) For the initialization error $\epsilon = \max\{|\hat{w}_1^{(0)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(0)} - \beta_{1,j}^*\|_2\}$:

$$\epsilon < \min\left\{\frac{\lambda_{\min}^3 w_{\min}^{*2}}{24\sqrt{10}c_2\lambda_{\max}^2 w_{\max}^{*2} r}, \frac{\lambda_{\min}^3 w_{\min}^{*3}}{4c_2\lambda_{\max} w_{\max}^{*3} r^2}, \frac{1}{6}\right\},$$

where c_1, c_2 is the same constants as in Assumption 1.

(d) The incoherence parameter $\xi = \max_{j\in[3], k\neq k'} |\langle \beta_{k,j}^*, \beta_{k',j}^*\rangle|$ satisfies:

$$\xi \leq \frac{\lambda_{\min}^3 w_{\min}^{*3}}{4c_2\lambda_{\max} w_{\max}^{*3} r^2}.$$

(e) The sample size n satisfies:

$$n \geq \min\left\{\frac{c_5\sigma^2 s^2\log(d)}{w_{\min}^{*2}p}, \frac{c_6\sigma s\log(d)\log(\sqrt{s^3/p})}{w_{\min}^* p}\right\},$$

where c_5 and c_6 are some positive constants.

Remarks:

1. The conditions are similar to corresponding conditions of Assumption 2 with an added affect of general rank r.

2. The condition (d) ensures control of correlations between the decomposed components across different ranks.

5.2 Main theorems

In this section, we introduce two theorems that show that estimation from Algorithm 1 gives precise results with a high probability. Moreover, we calculate the error of the estimator. Good performance from Algorithms 2 and 3 are crucial for a precise initialization for Algorithm 1 due to Assumption 2(c). Therefore, we present a theorem to theoretically prove that Algorithm 2 provides precise initialization with a high probability, and calculate its error. Zhou et al, 2021 [1] derive the non asymptotic error bound of the algorithms.

Theorem 5.1. Suppose that assumptions 1 and 2 hold. Then, for rank $r = 1$, the estimator from the t^{th} iteration of Algorithm 1 satisfies with high probability:

$$\max\{|\hat{w}_1^{(t)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(t)} - \beta_{1,j}^*\|_2\} \leq \underbrace{k^t \epsilon}_{\text{computational error}} + \underbrace{\frac{1}{1-k}\frac{C_1\sigma}{w_1^*}\sqrt{\frac{s\log(d)}{np}}}_{\text{statistical error}},$$

where $k = 6\sqrt{10}c_2\lambda_{\max}^2\epsilon/\lambda_{\min}^3 + 1/2 < 1$ is a positive coefficient, ϵ is defined in Assumption 2(c), c_2, q are defined in Assumption 1, $C_1 = (6\sqrt{10}\tilde{C}\lambda_{\max} + \tilde{C}_2\lambda_{\min}\sqrt{q})/\lambda_{\min}^2$ and $\tilde{C}, \tilde{C}_2$ are some positive constants.

The proof of this result is given in Appendix B.1.

Theorem 5.2. Suppose that assumptions 1 and 3 hold. Then, for a general rank r, the estimator from the t^{th} iteration of Algorithm 1 satisfies with high probability:

$$\max\{\max_k |\hat{w}_k^{(t)} - w_k^*|/w_k^*, \max_{k,j} \|\hat{\beta}_{k,j}^{(t)} - \beta_{k,j}^*\|_2\} \leq \underbrace{\tilde{k}^t \epsilon}_{\text{computational error}}$$

$$+ \underbrace{\frac{1}{1-\tilde{k}}\frac{C_1 w_{\max}^*\sigma}{w_{\min}^{*2}}\sqrt{\frac{s\log(d)}{np}}}_{\text{statistical error}},$$

where $\tilde{k} = \frac{6\sqrt{10}c_2\lambda_{\max}^2 w_{\max}^{*2} r}{\lambda_{\min}^3 w_{\min}^{*2}}\epsilon + \frac{c_1^2 c_2 \lambda_{\max} w_{\max}^{*3} r^2}{\lambda_{\min}^3 w_{\min}^{*3}}\epsilon + \frac{c_1^2 c_2 \lambda_{\max} w_{\max}^{*3} r^2}{\lambda_{\min}^3 w_{\min}^{*3}}\xi + \frac{1}{4} < 1$ is a positive coefficient, c_2 and q are defined in Assumption 1, $C_2 = (6\sqrt{10}\tilde{C}\lambda_{\max} + 12\tilde{C}_2\lambda_{\min}\sqrt{q})/\lambda_{\min}^2$ and $\tilde{C}, \tilde{C}_2$ are some positive constants.

The proof of this result is given in Appendix B.2.

Remark:

$\tilde{k}$ is greater than k, which indicates that the algorithm for the general case has a slower convergence rate. Moreover, $\tilde{k}$ increases with an increasing rank r. This can be expected since as the tensor estimation problem becomes more challenging, the algorithm will show a slower convergence rate.

Theorem 5.3. Suppose that Assumptions 1 and 2 (a, b, d) hold. Also, suppose that $L \geq C_1'$ for some large C_1', $|\sum_{i=1}^{n} n^{-1}\beta_{1,4}^{*T}x_i| \geq C_2'$ for some positive constant C_2'.

Then, the initial estimator produced by Algorithm 2 satisfies that

$$\max\{|\hat{w}_1^{(0)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(0)} - \beta_{1,j}^*\|_2\} = \mathcal{O}_p\left\{\sqrt{\frac{\log(d)}{nps^2}} + \frac{\sigma}{w_1*}\sqrt{\frac{s\log(d)}{np}}\right\}.$$

The proof of this result is given in Appendix B.3.

Remarks:

1. The result of the theorem shows that the initialization error decreases when n increases. Thus, the constant initialization error bound in Assumption 2(c) is guaranteed to hold as n increases.

2. The estimation error is slower than the statistical error rate in Theorem 5.1 when $\sigma/w_1^* \leq c/s^{1.5}$. This suggests that after obtaining the initial estimator using the spectral initialization algorithm, the alternating block algorithm 1 could further improve the error rate of the estimator.

6 Simulation studies and real dataset application

In this chapter, we aim to analyze the accuracy of the algorithms. Our goal is to investigate their performance using simulations and to analyze Parkinson's patience speech dataset.

6.1 Simulation

In this section, we perform some simulations to analyze the accuracy of the algorithm. Our goal is to investigate the performance of considered algorithms. Moreover, we want to explore the change in the error for different sample sizes, observation probability and fusion constraint.

To evaluate the performance of the estimator of the coefficient tensor B^*, we use the mean squared error (MSE) that is defined as:

$$\text{MSE} = \mathbb{E}\left[\left\|\hat{B} - B\right\|_F^2\right] = \sum_{i_1,i_2,i_3,i_4,i_5} \mathbb{E}\left[(\hat{B}_{i_1,i_2,i_3,i_4,i_5} - B_{i_1,i_2,i_3,i_4,i_5})^2\right]$$

$$= \mathbb{E}\left[\text{tr}(\text{vec}(B^* - \hat{B})(\text{vec}(B^* - \hat{B}))^T)\right].$$

For each considered set of model parameters $m = 30$ simulations are performed. The empirical mean squared error that is reported in the table is an average Error $= \frac{\sum_{l=1}^{m} \left\|\hat{B}_l - B\right\|_F^2}{m}$, where $\hat{B}_l$ is an estimate of the coefficient tensor in l^{th} simulation for set parameters. The standard error is also reported for the simulations.

The computational time of the algorithm is linear with the sample size and tensor dimension.

Two patterns of an observed data are considered: random missing and block missing.

6.1.1 Random missing

In this subsection, the missing data points are random, and don't follow a certain pattern. A fourth-order tensor response $Y_I \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times T}$ is generated as follows.

At the first step, the coefficient tensor is generated: $B^* \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times T \times q}$ as $B^* = \sum_{k=1}^{2} w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \beta_{k,3}^* \circ \beta_{k,4}^* \circ \beta_{k,5}^*$, where $d_1 = d_2 = d_3 = 32, T = 5, q = 5$ and the true rank $r = 2$. Entries of $\beta_{k,j}^*, j \in [4], k \in [2]$ are iid standard normal.

Then *Truncatefuse* operator is applied on $\beta_{k,j}^*, j \in [3], k \in [2]$ with the true sparsity and fusion parameters $(s_0 \times d_j, f_0 \times d_j), j \in [3], k \in [2]$, where $s_0 = 0.7$, and $f_0 = 0.7$ or $f_0 = 0.3$. Then, *Fuse* operator is applied to $\beta_{k,4}, k \in [2]$ with the true fusion parameter $f_0 \times T$. $\beta_{k,5}^*, k \in [2]$ is set as a vector of all ones: $\beta_{k,5}^* = (1, \ldots, 1)^T$. Then, each vector is normalized.

After those steps $\beta_{k,1}^*, \beta_{k,2}^*, \beta_{k,3}^*, \beta_{k,4}^*, \beta_{k,5}^*, k \in [2]$ meet the assumptions of the model. The weight is set as $w_k^* = 20, k \in [2]$.

Then the q-dimensional predictor vector x_i is generated such that its entries are iid normal with mean 2 and standard deviation 3. The error tensor $\mathcal{E}$ whose entries are iid standard normal is generated as well.

Next, the response tensor Y_i is computed following the model in (1).

Eight sets for simulations are performed, based on the probability of observed data: $p = 0.3$, $p = 0.7$, fusion constant: $f_0 = 0.3$, $f_0 = 0.7$, and sample size $n = 80$, $n = 150$.

Table 1: Simulation for random missing with $p = 0.3$

(a) $f_0 = 0.3$

n	Error	SE
80	0.0149	0.0017
150	0.0069	0.0003

(b) $f_0 = 0.7$

n	Error	SE
80	0.0228	0.0022
150	0.0112	0.0005

Table 2: Simulation for random missing with $p = 0.5$

(a) $f_0 = 0.3$

n	Error	SE
80	0.0062	0.0009
150	0.0028	0.0001

(b) $f_0 = 0.7$

n	Error	SE
80	0.0088	0.0012
150	0.0071	0.0003

As seen by the results, the error decreases when the observation probability (p) and sample size increase. This is consistent with the theoretical results. Additionally, incorporating the fusion structure improves the estimation accuracy, however, estimation error increases when the fusion constant increases.

6.1.2 Block missing

In this subsection, we consider the scenario where the unobserved data is located in blocks. In real life applications, this is a common situation. For example, a missing MRI scan would be a missing block for a subject at a certain time.

Two probability variables are introduced: p_n - probability that each subject has missing values and p_t - proportion of missing blocks for the subject by the time variable. For example, if there are 100 subjects and $T = 5, p_n = 0.8, p_t = 0.4$, then it means that $0.8 \times 100 = 80$ subjects have partially observed tensor, and for each of those 80 subjects, observations are missing 2 out of 5 times.

The simulation process stays the same as in the previous case with random missing.

Table 3: Simulation for block missing with $p_n = 0.8, p_t = 0.4$

(a) $f_0 = 0.3$

n	Error	SE
80	0.0067	0.0012
150	0.0044	0.0004

(b) $f_0 = 0.7$

n	Error	SE
80	0.0166	0.0022
150	0.0121	0.0007

Table 4: Simulation for block missing with $p_n = 0.8, p_t = 0.6$

(a) $f_0 = 0.3$

n	Error	SE
80	0.0151	0.0029
150	0.0085	0.0013

(b) $f_0 = 0.7$

n	Error	SE
80	0.0243	0.0033
150	0.0180	0.0019

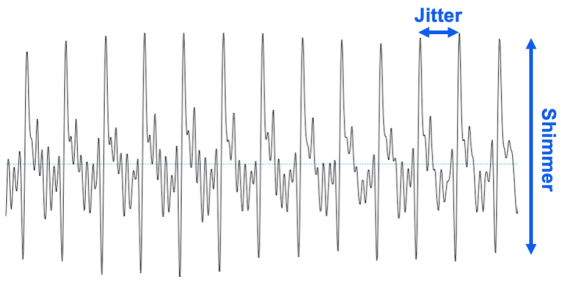
As with random missing, the error decreases when the observation probability (p) and sample size increase.

6.2 Analysis of a real dataset

In this section, we apply considered algorithms to a real-life dataset. The algorithm is illustrated by applying it to a voice analysis dataset of Parkinson's disease patients. Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves, which causes unintended or uncontrollable movements, such as shaking, stiffness, and difficulty with balance and coordination. There are currently no blood or laboratory tests to diagnose most cases of Parkinson's. Therefore, it is important to detect patterns that can help an early diagnosis of Parkinson's. One of the considered symptoms of the disease is that patient's speech becomes soft or slurred.

The analyzed data is taken from the study on telemonitoring of Parkinson's disease progression by non-invasive speech tests Tsanas et al, 2009 [8]. It is of interest to see how different speech attributes relate to the disease's progression. To measure the Parkinson's progression, Unified Parkinson's disease rating scale (UPDRS) is used.

The speech data was collected over 6 months from $n = 42$ participants . Each months, there were several tests taken. For every voice recording, there are a number of attributes that were measured. Based on previous research and simplicity of interpretation, those three attributes are chosen: Jitter.PPQ5, Shimmer.APQ5 and PPE.



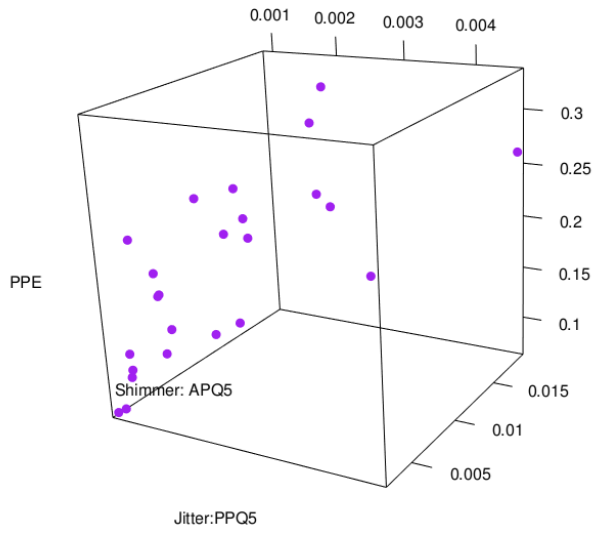
Jitter and shimmer are acoustic characteristics of voice signals, and they are caused by irregular vocal fold vibration. They are perceived as roughness, breathiness, or hoarseness in a speaker's voice. PPQ5 is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period. Let

$$\text{PPQ5} = \frac{\sum_{i=3}^{N-2} \left| T_i - \frac{T_{i-2}+T_{i-1}+T_i+T_{i+1}+T_{i+2}}{5} \right| /(N-4)}{\sum_{i=1}^{N} T_i/N}.$$

APQ5 is the five-point Amplitude of Perturbation Quotient, the average absolute difference between the amplitude of a period and the mean amplitudes of it and its four closest neighbors, divided by the average amplitude. Moreover, PPE - Pitch Period Entropy is a nonlinear measure of fundamental frequency variation.

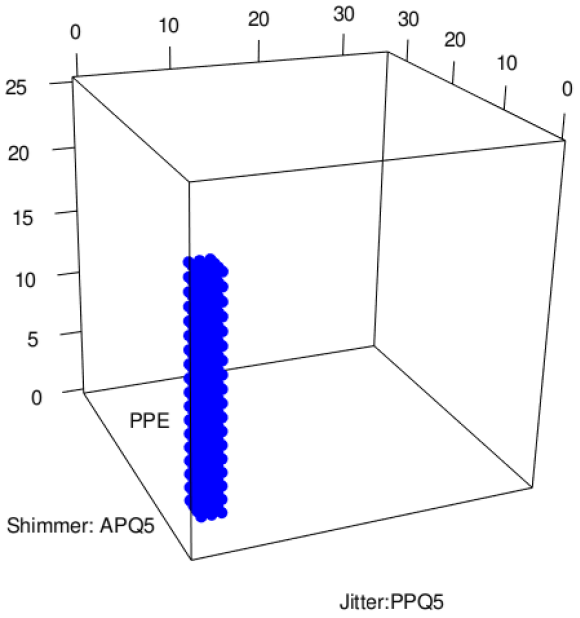
Each group of speech tests after preprocessing and mapping is summarized in the form of $35 \times 34 \times 25$ tensor. For each participant, there are 6 tensors for every month. For each subject, we stack those tensors collected over time as a fourth-order tensor, which is to serve as the response tensor Y_i with dimensions $35 \times 34 \times 25 \times 6$. 20% of the data is missing in blocks: for example, if a subject wouldn't record their test in a month.

Figure 1: For each subject and for each month of the study, we have a group of voice recordings. Each voice recording is the point in a 3D space with coordinates as its corresponding attributes. As picture shows, this process would create sparse tensors.



The predictor x_i consists of a continuous variable Unified Parkinson's disease rating scale, age and sex. The goal is to identify the values or a pattern of Jitter.PPQ5, Shimmer.APQ5 and PPE that relates to progression of Parkinson's.

Figure 2: The estimated regression coefficient tensor after applying proposed algorithms. The points show non-zero entries of the tensor.



As a result, the estimation of the regression coefficient tensor was obtained. It is seen that the estimate identifies that the relationships between Jitter.PPQ5 and UPDRS, as well as, Shimmer.APQ5 and UPDRS are mostly significant for lower values of Jitter.PPQ5 and Shimmer.APQ5. In regards of PPE, it is a highly significant variable for predicting the progression of Parkinson's disease. PPE indicates impaired pitch control that could be interpreted as deteriorating muscle coordination. Those findings are consistent with existing research on speech analysis for Parkinson's diagnosis Little et al, 2011 [12], Tsanas et al, 2009 [8].

7 Conclusion

In this major paper, we study with more details the alternating updating algorithm proposed by Zhou et al, 2021 [1]. This proposed method is unique in terms of estimation algorithm, theoretical properties, and regularity conditions. Zhou et al, 2021 [1] developed an efficient algorithm that deals with a challenge of unobserved tensor data. Without completing the data, the algorithms estimates solution of a non-convex optimization problem. The non-convexity causes the theoretical explanation to be highly nontrivial.

After careful consideration of the proofs, this analysis shows that one of the assumptions could be dropped. Specifically, we weaken assumption on the bound of tensor coefficient's norm. Thus, those findings show that the algorithm could be applied for a more general model. Moreover, we calculated the estimation error and proved that the considered algorithm gives a precise estimator with a high probability. Also, we analyzed an initialization algorithms which is crucial to the good performance of the estimation algorithm. The efficiency and accuracy of the algorithm were illustrated using simulations. Two data patterns - block and random missing were considered, and both showed that the estimation error decreases when the observation probability and sample size increase. Theoretical analysis proves the same result. The computational time of the algorithm is linear with the sample size and tensor dimension. The method was also applied to a speech data of Parkinson's patients. As a results, an important pattern of changes in the speech attributed were discovered. Pitch Period Entropy showed strong significance for Parkinson's diagnosis. This finding is consistent with the results in Little et al, 2011 [12], where the analysis proved that PPE has the best classification performance out of all considered variables. Our findings are also consistent with the results given in Tsanas et al, 2009 [8]. More precisely, our analysis highlights a significance for Shimmer, Jitter and PPE.

References

[1] J. Zhou, W.W. Sun, J. Zhang, L. Li (2021): "Partially Observed Dynamic Tensor Response Regression, Journal of the American Statistical Association", DOI: 10.1080/01621459.2021.1938082.

[2] H. Zhou, L. Li, H. Zhu (2013): "Tensor Regression with Application in Neuroimaging Data Analysis", Journal of the American Statistical Association, 108:502, 540-552, DOI: 10.1080/01621459.2013.776499.

[3] W. W. Sun, L. Li (2017): "Store: Sparse Tensor Response Regression and Neuroimaging Analysis", Journal of Machine Learning Research, 18, 1–37.

[4] T. G. Kolda (2006): "Multilinear Operators For Higher-Order Decompositions", Sandia National Laboratories, Tech. Rep. No. SAND2006-2081.

[5] T. G. Kolda, B. W. Bader (2009): "Tensor Decompositions and Applications", SIAM Review, 51, 455–500.

[6] K.-H. Thung, C.-Y. Wee, P.-T. Yap, D. Shen (2016): "Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans", Brain Struct Funct 221, 3979–3995.

[7] ADHD-200 Sample Initiative
($http://fcon\_1000.projects.nitrc.org/indi/adhd200/$)

[8] A. Tsanas, M. A. Little, P. E. McSharry, Lorraince O. Ramig (2009): "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests", IEEE Transactions on Biomedical Engineering, vol. 57, no. 4, pp. 884-893.

[9] N.I. Bruce, B. Murthi, R.C. Rao (2017) : "A dynamic model for digital advertising: the effects of creative format, message content and targeting on engagement, Journal of Marketing Research", 54(2), 202–218.

[10] P. Jain and S. Oh (2014): "Provable Tensor Factorization with missing data", MIT Press, Volume 1 (NIPS'14), 1431–1439.

[11] D.Xia and M. Yuan (2019): "On polynomial time methods for exact low rank tensor completion", Found Comput Math 19, 1265–1313.

[12] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig (2011): "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", IEEE transactions on biomedical engineering 56(4):1015.

[13] D. Xia, M.Yuan, C.Zhang (2020): "Statistically optimal and computationally efficient low rank tensor completion from noisy entries", Annals of Statistics, 49(1), 76–99.

[14] T. Ryota, S. Taiji (2014): "Spectral norm of random tensors", arXiv:1407.1870.

[15] A. Zhang (2019): "Cross efficient low-rank tensor completion", Annals of Statistics, 47, 936–964.

[16] X.-T. Yuan and T. Zhang (2013): "Truncated power method for sparse eigenvalue problems", Journal of Machine Learning Research, 14, 899– 925.

[17] W. Sun, J. Lu and G.Cheng (2017): "Provable sparse tensor decomposition", Journal of the Royal Statistical Society, Series B, 79, 899– 916.

[18] C.Cai, G.Li, H. Poor and Y.Chen (2021): "Nonconvex low-rank tensor completion from noisy data", NeurIPS, 32, 1863–1874.

[19] G.W.Stewart and J.Sun (1990): "Matrix perturbation theory", Academic Press.

[20] G.Bennett (1962): "Probability inequalities for the sum of independent random variables", J. Amer. Statist. Assoc. 57 33-45.

[21] J.M.Kohler and A. Lucchi (2017): "Sub-sampled Cubic Regularization for Non-convex Optimization", arXiv:1705.05933.

[22] J.A.Tropp (2015): "An Introduction to Matrix Concentration Inequalities", arXiv:1501.01571.

Appendix A Some useful preliminary results

This Appendix contains propositions and their proofs, as well as theorems and lemmas, that are used to prove Theorems 5.1 - 5.3.

Theorem A.1 (**Bernstein's inequality**)**.** Let $X_1, X_2, \ldots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely for all $i \in [n]$. Then, for all positive t:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum\limits_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{1}{3}Mt}\right)$$

The proof of this theorem is given in Bennett, 1962 [20].

Theorem A.2 (**Vector Bernstein's inequality**)**.** Let $X_1, \ldots X_n$ be independent vector-valued random variables with common dimension d and assume that

$$\mathbb{E}(X_i) = 0, \|X_i\|_2 \leq \mu, \mathbb{E}\left[\|X_i\|^2\right] \leq \sigma^2$$

Then

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\| \geq \epsilon\right) \leq \exp\left(\frac{1}{4} - \frac{n\epsilon^2}{8\sigma^2}\right)$$

The proof of this theorem is given in Kohler and Lucchi, 2017 [21].

Theorem A.3 (**Matrix Bernstein's inequality**)**.** Let $X_1, \ldots X_n$ be independent $d_1 \times d_2$ random matrices and assume that $\mathbb{E}(X_i) = 0, \|X_i\| \leq B$. Define $W = \sum_{i=1}^{n} X_i$ and

$$\delta^2 := \max\left\{\left\|\mathbb{E}WW^T\right\|, \left\|\mathbb{E}W^TW\right\|\right\}.$$

Then

$$\mathbb{P}\left(\|W\| \geq t\right) \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{\sigma^2 + Bt/3}\right).$$

The proof of this theorem is given in Tropp, 2015 [22].

Theorem A.4 (**Wedin's theorem**)**.** (Theorem 4.4 from Stewart and Sun, 1990 [19])
Let $A, E \in \mathbb{R}^{m\times n}$ with $m \geq n$. Suppose that A has singular value decomposition.

$$\begin{bmatrix} U_1^T \\ U_2^T \\ U_3^T \end{bmatrix} A \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix}.$$

Let $\tilde{A} = A+E$, with analogous singular value decomposition $(\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{V}_1, \tilde{V}_2, \tilde{\Sigma}_1, \tilde{\Sigma}_2)$. Let $\delta > 0$, $\delta = \min\left\{\min_{i,j}|\Sigma_1[i,i] - \Sigma_2[j,j]|, \min_i \Sigma_1[i,i]\right\}$. If $\delta \geq 4\|E\|_2$, then the distance between U and $\tilde{U}$ is bounded by $\mathcal{O}(\|E\|_2/\delta)$.

Lemma A.1. (Lemma 5 from Xia et al, 2020 [13].)

Let $X_1, \ldots, X_n \in \mathbb{R}^{m_1 \times m_2}$ be random matrices with zero mean. Suppose that $\max_{1 \le i \le n} \| \|X_i\| \|_{\psi_\alpha} \le U^{(\alpha)} < \infty$ for some $\alpha \ge 1$. Let

$$\delta^2 := \max \left\{ \left\| \sum_{i=1}^n \mathbb{E} X_i X_i^T \right\|, \left\| \sum_{i=1}^n \mathbb{E} X_i^T X_i \right\| \right\}.$$

Then there exist a universal constant $C > 1$ such that for all $t > 0$, the following bound holds with probability at least $1 - e^{-t}$,

$$\left\| \frac{X_1 + \cdots + X_n}{n} \right\| \le C \max \Bigg\{ \delta \frac{\sqrt{t + \log(m_1 + m_2)}}{n},$$

$$U^{(\alpha)} \left(\log \frac{\sqrt{n} U^{(\alpha)}}{\delta} \right) \frac{t + \log(m_1 + m_2)}{n} \Bigg\}.$$

Lemma A.2. (Theorem 1 from Ryota and Taiji, 2014 [14].)

Let $X \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is a K-way tensor. The spectral norm of X is defined as follows:

$$|||X||| = \sup_{u_1, u_2, \ldots, u_k} X(u_1, u_2, \ldots, u_k), u_k \in S_{n_k - 1}, k = 1, \ldots, K$$

where $X(u_1, \ldots, u_k) = \sum_{i_1, i_2, \ldots, i_k} X_{i_1, i_2, \ldots, i_k} u_{1 i_1} \ldots u_{K i_K}$ and $S_{n_k - 1}$ is the unit sphere in $\mathbb{R}^{n_k}$. Then,

$$\mathbb{P}(\| |X| \| \ge t) \le \sum_{\bar{u}_1 \in C_1, \ldots, \bar{u}_K \in C_K} \mathbb{P} \left(X(\bar{u}_1, \ldots, \bar{u}_K) \ge \frac{t}{2} \right)$$

Lemma A.3. (Lemma 4 from Zhang, 2019 [15].)

For a scalar a^* and any sequence $(a_1, a_2, \ldots, a_S)$,

$$\sum_{i=1}^S (a_i - a^*)^2 \ge S \left(S^{-1} \sum_{i=1}^S a_i - a^* \right)^2,$$

where the equality holds if and only if $a_1 = a_2 = \cdots = a_S$.

Lemma A.4. (Lemma D.6 from Cai et al, 2021 [18].)

Let U and V be two $d \times r$ matrices, each with orthogonal columns. Suppose that $\|UU^T - VV^T\| \le \delta$. Then, for any unit vector $u_o \in \mathbb{R}^d$ lying in span$U$,we have

$$\|P_V(u_0)\| \ge \sqrt{1 - \delta^2} \text{ and } \|P_{V\perp}(u_0)\| \le \delta,$$

where $P_V(u_0) = VV^T u_0$.

Proposition A.1. Let $f(x) = \sum_i \alpha_i \sum_j \left(a_{i,j} - b_{i,j}x\right)^2, x \in \mathbb{R}$, for given $\alpha_i \geq 0$, $a_{i,j} \in \mathbb{R}, b_{i,j} \in \mathbb{R}$ for all $i = 1, \ldots, n$, $j = 1, \ldots, m$. Then, $f(x)$ reaches its minimum at

$$x_{\min} = \frac{\sum_i \alpha_i \sum_j a_{i,j} b_{i,j}}{\sum_i \alpha_i \sum_j b_{i,j}^2}.$$

Proof. We have

$$\sum_i \alpha_i \sum_j (a_{i,j} - b_{i,j}x)^2 = \sum_i \alpha_i \sum_j a_{i,j}^2 + \sum_i \alpha_i \sum_j (-2a_{i,j}b_{i,j}x) + \sum_i \alpha_i \sum_j b_{i,j}^2 x^2.$$

Since x does not depend on i, j, we have

$$\sum_i \alpha_i \sum_j \left(a_{i,j} - b_{i,j}x\right)^2 = \sum_i \alpha_i \sum_j a_{i,j}^2 - 2x \sum_i \alpha_i \sum_j a_{i,j}b_{i,j} + x^2 \sum_i \alpha_i \sum_j b_{i,j}^2.$$

Let $D = \sum_i \alpha_i \sum_j b_{i,j}^2 \geq 0$, $F = \sum_i \alpha_i \sum_j a_{i,j}b_{i,j}$ and $C = \sum_i \alpha_i \sum_j a_{i,j}^2$. Then, we have

$$f(x) = Dx^2 - 2xF + C, x \in \mathbb{R} \text{ for given } D \geq 0.$$

Hence, we solve a quadratic minimization problem with positive leading coefficient. Then, we have

$$x_{\min} = \frac{2F}{2D} = \frac{\sum_i \alpha_i \sum_j a_{i,j}b_{i,j}}{\sum_i \alpha_i \sum_j b_{i,j}^2}.$$

□

Proposition A.2. Let x, y be two q-column unit vectors. Then

$$\langle x, y \rangle = 1 - \frac{\|x - y\|^2}{2}.$$

Proof. We have

$$\langle x, y \rangle = \langle (x - y) + y, (y - x) + x \rangle.$$

Recall that $\|x\|^2 = \langle x, x \rangle$ for any vector x, then

$$\langle x, y \rangle = -\|x-y\|^2 + \langle x-y, x \rangle + \langle y, y-x \rangle + \langle y, x \rangle = -\|x-y\|^2 + 1 - \langle y, x \rangle + 1 - \langle y, x \rangle$$

$$+\langle y, x \rangle = 2 - \|x - y\|^2 - \langle y, x \rangle.$$

Hence, we have

$$\langle x, y \rangle = 1 - \frac{\|x - y\|^2}{2}.$$

□

Proposition A.3. Let A_1 and A_2 be two positive defined matrices, such that $1 - \|A_1 - A_2\|\|A_2^{-1}\| > 0$. Then

$$\|A_1^{-1}\| \leq \frac{\|A_2^{-1}\|}{1 - \|A_1 - A_2\|\|A_2^{-1}\|}.$$

Proof. Since A_1 and A_2 are positive definite matrices, A_1, A_2 are invertible, and A_1^{-1}, A_2^{-1} are positive definite. Using matrix norm properties, we have

$$\|A_1 - A_2\|\|A_1^{-1}\|\|A_2^{-1}\| \geq \|(A_1 - A_2)A_1^{-1}A_2^{-1}\| = \|A_1^{-1} - A_2^{-1}\|.$$

Recall that $\|A_1^{-1} - A_2^{-1}\| \geq \|A_1^{-1}\| - \|A_2^{-1}\|$. Therefore,

$$\|A_1 - A_2\|\|A_1^{-1}\|\|A_2^{-1}\| \geq \|A_1^{-1}\| - \|A_2^{-1}\|.$$

Then,

$$\|A_1^{-1}\| - \|A_1 - A_2\|\|A_1^{-1}\|\|A_2^{-1}\| \leq \|A_2^{-1}\|.$$

Recall that $1 - \|A_1 - A_2\|\|A_2^{-1}\| > 0$. Then

$$\frac{\|A_1^{-1}\| - \|A_1 - A_2\|\|A_1^{-1}\|\|A_2^{-1}\|}{1 - \|A_1 - A_2\|\|A_2^{-1}\|} \leq \frac{\|A_2^{-1}\|}{1 - \|A_1 - A_2\|\|A_2^{-1}\|}$$

$$\frac{\|A_1^{-1}\|(1 - \|A_1 - A_2\|\|A_2^{-1}\|)}{1 - \|A_1 - A_2\|\|A_2^{-1}\|} \leq \frac{\|A_2^{-1}\|}{1 - \|A_1 - A_2\|\|A_2^{-1}\|}.$$

Therefore,

$$\|A_1^{-1}\| \leq \frac{\|A_2^{-1}\|}{1 - \|A_1 - A_2\|\|A_2^{-1}\|}.$$

□

Proposition A.4. Let X be a sub-Gaussian random variable with variance σ^2. Then, for any positive integer $k \geq 1$,

$$\mathbb{E}(|X|^k) \leq (2\sigma^2)^{k/2} k\Gamma(k/2).$$

Proof.

$$\mathbb{E}(|X|^k) = \int_0^\infty \mathbb{P}(|X|^k > t)\, dt = \int_0^\infty \mathbb{P}(|X| > t^{1/k})\, dt.$$

Since X be a sub-Gaussian random variable with variance σ^2, we have that $\mathbb{P}(|X| > t) \leq 2\exp\left(\frac{-t^2}{2\sigma^2}\right)$. Therefore, we get

$$\mathbb{E}(|X|^k) \leq \int_0^\infty 2\exp\left(\frac{-t^{2/k}}{2\sigma^2}\right)\, dt.$$

Let $u = \frac{t^{2/k}}{2\sigma^2}$. Then, $t = (2\sigma^2 u)^{k/2}$ and

$$\int_0^\infty 2\exp\left(\frac{-t^{2/k}}{2\sigma^2}\right) dt = \int_0^\infty 2e^{-u}(2\sigma^2)^{k/2}\frac{k}{2}e^{-u}u^{k/2-1}\,du = (2\sigma^2)^{k/2}k\int_0^\infty e^{-u}u^{k/2-1}\,du$$

$$= (2\sigma^2)^{k/2}k\Gamma(k/2).$$

Therefore, we have that

$$\mathbb{E}(|X|^k) \leq (2\sigma^2)^{k/2}k\Gamma(k/2).$$

□

Corollary A.1. Let δ be a random indicator variable, such that $\delta = 1$ with probability p and $\delta = 0$ with probability $1-p$. Let $\mathcal{E}$ be a sub-Gaussian random variable with variance σ^2. Then,

$$\mathbb{E}(\delta\mathcal{E}^2) \leq 4\sigma^2\sqrt{p}.$$

Proof. By Hölder's inequality, we have

$$\mathbb{E}(\delta\mathcal{E}^2) \leq \sqrt{\mathbb{E}\delta^2}\sqrt{\mathbb{E}\mathcal{E}^4}.$$

By Proposition A.4, we have that for any positive integer $k \geq 1$, $\mathbb{E}(|\mathcal{E}|^k) \leq (2\sigma^2)^{k/2}k\Gamma(k/2)$. Then, we get

$$\mathbb{E}(\mathcal{E}^4) \leq (2\sigma^2)^2 4\Gamma(2) = 16\sigma^4.$$

Moreover, it is clear that $\mathbb{E}\delta^2 = 1^2 p + 0(1-p) = p$. Therefore,

$$\mathbb{E}(\delta\mathcal{E}^2) \leq \sqrt{p}4\sigma^2.$$

□

Proposition A.5. Suppose that the conditions of Theorem 5.1 hold and let

$$I_1 = \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\left\{\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle - 1\right\}\beta_{1,3}^*.$$

Then $\|I_1\| \leq \frac{2\lambda_{\max}}{\lambda_{\min}}\epsilon^2$.

Proof. By norm properties, we have

$$\|I_1\| = \left|\frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|\left|\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle - 1\right|\|\beta_{1,3}^*\|.$$

Further, by combining the fact that $||x| - |y|| \leq |x - y|$ and Assumption 2(c), we have

$$||\hat{w}_1| - |w_1^*|| \leq |\hat{w}_1 - w_1^*| < \epsilon w_1^* < \frac{1}{2} w_1^*.$$

Then

$$-\frac{1}{2} w_1^* < |\hat{w}_1| - |w_1^*| \Rightarrow |\hat{w}_1| > \frac{1}{2} w_1^*. \tag{A.1}$$

Applying the Cauchy–Schwarz inequality, we get

$$\left| \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} \alpha_{i,1}^*}{n} \right| \leq \sqrt{\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}^2}{n}} \sqrt{\sum_{i=1}^{n} \frac{\alpha_{i,1}^{*2}}{n}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_{1,4}^T x_i x_i^T \hat{\beta}_{1,4}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \beta_{1,4}^{*T} x_i x_i^T \beta_{1,4}^*}$$

$$= \sqrt{\hat{\beta}_{1,4}^T \sum_{i=1}^{n} \frac{x_i x_i^T}{n} \hat{\beta}_{1,4}} \sqrt{\hat{\beta}_{1,4}^{*T} \sum_{i=1}^{n} \frac{x_i x_i^T}{n} \hat{\beta}_{1,4}^*}.$$

Note that if X is a symmetric matrix and v is a unit vector, then $v^T X v \leq \lambda_{\max}$, where $\lambda_{\max}$ is a maximum eigenvalue of X. Further, recall that $\sum_{i=1}^{n} \frac{x_i x_i^T}{n}$ is a symmetric matrix. Hence, using Assumption 1, we have

$$\left| \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} \alpha_{i,1}^*}{n} \right| \leq \sqrt{\lambda_{\max}} \sqrt{\lambda_{\max}} = \lambda_{\max}.$$

Likewise, we have that

$$\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}^2}{n} = \hat{\beta}_{1,4}^T \sum_{i=1}^{n} \frac{x_i x_i^T}{n} \hat{\beta}_{1,4} \geq \lambda_{\min} > 0.$$

Therefore,

$$\left| \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \right| \leq \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{A.2}$$

Then, we apply Proposition A.2:

$$\left| 1 - \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \right| = \left| 1 - \frac{1}{4} (2 - \|\hat{\beta}_{1,1} - \beta_{1,1}^*\|^2)(2 - \|\hat{\beta}_{1,2} - \beta_{1,2}^*\|^2) \right|$$

$$= \left| \frac{1}{2} \|\hat{\beta}_{1,1} - \beta_{1,1}^*\|^2 + \frac{1}{2} \|\hat{\beta}_{1,2} - \beta_{1,2}^*\|^2 - \frac{1}{4} \|\hat{\beta}_{1,1} - \beta_{1,1}^*\|^2 \|\hat{\beta}_{1,2} - \beta_{1,2}^*\|^2 \right|$$

$$\leq \frac{1}{2} \|\hat{\beta}_{1,1} - \beta_{1,1}^*\|^2 + \frac{1}{2} \|\hat{\beta}_{1,2} - \beta_{1,2}^*\|^2.$$

Since by Assumption 2(c), we have $\|\hat{\beta}_{1,1} - \beta^*_{1,1}\| \leq \epsilon$, we get that $\|\hat{\beta}_{1,1} - \beta^*_{1,1}\|^2 \leq \epsilon^2$ and $\|\hat{\beta}_{1,2} - \beta^*_{1,2}\|^2 \leq \epsilon^2$. Hence, we have

$$\left|1 - \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle\right| \leq \epsilon^2. \tag{A.3}$$

Hence, by combining (A.1), (A.2), (A.3) and the fact that $\|\beta^*_{1,3}\| = 1$, we get

$$\|I_1\| \leq \frac{2\lambda_{\max}}{\lambda_{\min}}\epsilon^2.$$

□

Proposition A.6. Suppose that the conditions of Theorem 5.1 hold and let

$$J_1 = \left| \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \hat{\alpha}_{i,1} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2} - \alpha^*_{i,1}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2} \right|^2.$$

Then,

$$J_1 \leq \left\{ 4\frac{\lambda^2_{\max}}{\lambda^2_{\min}} + 1 \right\} c_1^2\epsilon^2.$$

Proof.

$$J_1 = \left| \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \hat{\alpha}_{i,1} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2} - \alpha^*_{i,1}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2} \right|^2$$

$$= \left(\frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \hat{\alpha}_{i,1} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2} \right)^2$$

$$-2\frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \hat{\alpha}_{i,1}\alpha^*_{i,1} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2} + \alpha^{*2}_{i,1}\beta^{*2}_{1,1,l_1}\beta^{*2}_{1,2,l_2}.$$

Since β's are unit vectors, $|\beta_{k,j,i}| \leq 1$. Then, by Cauchy–Schwarz inequality $|\langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle| \leq \|\beta^*_{1,1}\|\|\hat{\beta}_{1,1}\| \leq 1$ and $\langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle^2 \leq |\langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle|$. Therefore, by combining with proven in Zhou et al, 2021 [1], we have

$$J_1 \leq \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle^2 \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle^2 \left[\left\{ \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \right\}^2 \hat{\alpha}^2_{i,1} - 2\frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^{n} \hat{\alpha}^2_{i,1}/n} \hat{\alpha}_{i,1}\alpha^*_{i,1} \right] + \alpha^{*2}_{i,1}.$$

This gives

$$J_1 \leq \langle \beta^*_{1,1}, \hat{\beta}_{1,1} \rangle^2 \langle \beta^*_{1,2}, \hat{\beta}_{1,2} \rangle^2 \Bigg[\left\{ \frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \right\}^2 \left\{ \hat{\alpha}_{i,1} - \underbrace{\alpha^*_{i,1} + \alpha^*_{i,1}}_{=0} \right\}^2 + \underbrace{\alpha^{*2}_{i,1}}_{\geq 0}$$

$$-2 \frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \left\{ \hat{\alpha}_{i,1} - \alpha^*_{i,1} + \alpha^*_{i,1} \right\} \alpha^*_{i,1} \Bigg] + \alpha^{*2}_{i,1} \underbrace{(1 - \langle \beta^*_{1,1}, \hat{\beta}_{1,1} \rangle^2 \langle \beta^*_{1,2}, \hat{\beta}_{1,2} \rangle^2)}_{\geq 0}.$$

Since $\langle \beta^*_{1,1}, \hat{\beta}_{1,1} \rangle^2 \leq 1$ and $\langle \beta^*_{1,2}, \hat{\beta}_{1,2} \rangle^2 \leq 1$, we have

$$J_1 \leq \left[\frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \left\{ \hat{\alpha}_{i,1} - \alpha^*_{i,1} + \alpha^*_{i,1} \right\} - \alpha^*_{i,1} \right]^2 + \alpha^{*2}_{i,1} (1 - \langle \beta^*_{1,1}, \hat{\beta}_{1,1} \rangle^2 \langle \beta^*_{1,2}, \hat{\beta}_{1,2} \rangle^2).$$

Then, by (A.3) and the fact that $\alpha^{*2}_{i,1} \leq c_1^2$ from Assumption 1, we have

$$J_1 \leq \left[\frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \left\{ \hat{\alpha}_{i,1} - \alpha^*_{i,1} \right\} + \alpha^*_{i,1} \frac{\sum_{i=1}^n \hat{\alpha}_{i,1} (\alpha^*_{i,1} - \hat{\alpha}_{i,1})/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \right]^2 + c_1^2 \epsilon^2.$$

Therefore,

$$J_1 \leq \left[\frac{\sum_{i=1}^n \hat{\alpha}_{i,1} (\alpha^*_{i,1} - \hat{\alpha}_{i,1})/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \right]^2 \alpha^{*2}_{i,1} + 2(\hat{\alpha}_{i,1} - \alpha^*_{i,1}) \alpha^*_{i,1} \frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \frac{\sum_{i=1}^n \hat{\alpha}_{i,1} (\alpha^*_{i,1} - \hat{\alpha}_{i,1})/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n}$$

$$+ \left[\frac{\sum_{i=1}^n \hat{\alpha}_{i,1} \alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \right]^2 \{ \hat{\alpha}_{i,1} - \alpha^*_{i,1} \}^2 + c_1^2 \epsilon^2.$$

Using Assumption 1, i.e. $\|x_i\| \leq c_1, \|\hat{\beta}_{i,1} - \beta^*_{i,1}\| < \epsilon$, and (A.2), we get:

$$J_1 \leq \frac{\lambda^2_{\max}}{\lambda^2_{\min}} c_1^2 \epsilon^2 + 2 \frac{\lambda^2_{\max}}{\lambda^2_{\min}} c_1^2 \epsilon^2 + \frac{\lambda^2_{\max}}{\lambda^2_{\min}} c_1^2 \epsilon^2 + c_1^2 \epsilon^2 = \left\{ 4 \frac{\lambda^2_{\max}}{\lambda^2_{\min}} + 1 \right\} c_1^2 \epsilon^2.$$

□

Proposition A.7. Suppose that the conditions of Theorem 5.1 hold and let

$$J_2 = \frac{1}{p}\sum_{i,l_1,l_2}\frac{1}{n^2}\hat{\beta}^2_{1,1,l_1}\hat{\beta}^2_{1,2,l_2}\beta^{*2}_{1,3,l}\hat{\alpha}^2_{i,1}\left\{\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n}\hat{\alpha}_{i,1}\langle\beta^*_{1,1},\hat{\beta}_{1,1}\rangle\langle\beta^*_{1,2},\hat{\beta}_{1,2}\rangle\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right.$$

$$\left.-\alpha^*_{i,1}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2}\right\}^2.$$

Then,

$$J_2 \leq \frac{\lambda_{\max}}{np}\frac{\mu^3}{s^{1.5}}\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2\epsilon^2.$$

Proof. Note that

$$J_2 = \frac{1}{p}\sum_{i,l_1,l_2}\frac{1}{n^2}\hat{\beta}^2_{1,1,l_1}\hat{\beta}^2_{1,2,l_2}\beta^{*2}_{1,3,l}\hat{\alpha}^2_{i,1}J_1,$$

where J_1 is defined in Proposition A.6. Since $\beta^*_{1,1,l_1} \leq \frac{\mu}{\sqrt{s}}$ for all $l_1 \in [d_1]$, $\beta^*_{1,2,l_2} \leq \frac{\mu}{\sqrt{s}}$ for all $l_2 \in [d_2]$, we have

$$J_2 \leq \frac{1}{np}\sum_i \frac{\hat{\alpha}^2_{i,1}}{n}\frac{\mu^4}{s^2}\sum_l \beta^{*2}_{1,3,l}J_1.$$

Hence, by applying Proposition A.6, we get

$$J_2 \leq \frac{\lambda_{\max}}{np}\frac{\mu^3}{s^{1.5}}\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2\epsilon^2.$$

□

Proposition A.8. Suppose that the conditions of Theorem 5.1 hold and let

$$II_1 = \frac{w^*_1}{\hat{w}_1}A^{-1}\left\{B - A\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n}\langle\beta^*_{1,1},\hat{\beta}_{1,1}\rangle\langle\beta^*_{1,2},\hat{\beta}_{1,2}\rangle\right\}\beta^*_{1,3},$$

where A and B are diagonal matrices with diagonal entry,

$$A_{ll} = \sum_{i=1}^n \hat{\alpha}^2_{i,1}/n\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}^2_{1,1,l_1}\hat{\beta}^2_{1,2,l_2}$$

$$B_{ll} = \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2}. \tag{A.4}$$

Then, with probability at least $1 - e^{\frac{1}{4}}/d^{10}$, we have $\|II_1\| \leq \frac{4\gamma\epsilon}{\lambda_{\min}}$.

Proof. We have

$$II_1 = \frac{w_1^*}{\hat{w}_1}\left\{\frac{1}{p}A\right\}^{-1}\frac{1}{p}\left\{B - A\frac{\sum_{i=1}^{n}\hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^{n}\hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*,\hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*,\hat{\beta}_{1,2}\rangle\right\}\beta_{1,3}^*.$$

We first find an upper bound of each diagonal entry of the matrix $p^{-1}A$. Assume that for all $i \in [n], l_1 \in [d_1], l_2 \in [d_2], l \in [d_3]$, $\delta_{i,l_1,l_2,l}$ is independent with $\hat{\alpha}_{i,1}, \hat{\beta}_{1,1,l_1}, \hat{\beta}_{1,2,l_2}$. Let $Z_{i,l_1,l_2} = p^{-1}n^{-1}\hat{\alpha}_{i,1}^2\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2$. Then $p^{-1}A$ has the form of

$$\frac{1}{p}\sum_{i=1}^{n}\hat{\alpha}_{i,1}^2/n\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2 = \sum_{i,l_1,l_2}Z_{i,l_1,l_2}.$$

Note that,

$$\mathbb{E}(\delta_{i,l_1,l_2,l}) = 1\times p + 0\times(1-p) = p,$$
$$\mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}) = \hat{\alpha}_{i,1}^2/n\sum_{l_1,l_2}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2.$$

Hence, since $|\hat{\alpha}_{i,1}| = |\hat{\beta}_{1,4}^T x_i| \le c_1$ and $\max_{l\in d_j}|\beta_{k,j,l}^*| \le \mu/\sqrt{s}$ from Assumption 1, we have:

$$\left|Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right| \le \left|\left(\frac{1}{p}\delta_{i,l_1,l_2,l} - 1\right)\frac{1}{n}\right| c_1^2\frac{\mu^4}{s^2} \le \frac{c_1^2\mu^4}{nps^2}.$$

Then,

$$\begin{aligned}
&\sum_{i,l_1,l_2}\mathbb{E}\left(\left[Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right]^2\Big|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right)\\
&= \sum_{i,l_1,l_2}\mathbb{E}\left(Z_{i,l_1,l_2}^2|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right) - \mathbb{E}^2\left(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right)\\
&= \left(\frac{1}{p} - 1\right)\sum_{i,l_1,l_2}\frac{1}{n^2}\hat{\alpha}_{i,1}^4\hat{\beta}_{1,1,l_1}^4\hat{\beta}_{1,2,l_2}^4 \le \frac{1}{p}\sum_{i,l_1,l_2}\frac{1}{n^2}\hat{\alpha}_{i,1}^4\hat{\beta}_{1,1,l_1}^4\hat{\beta}_{1,2,l_2}^4,
\end{aligned}$$

and then

$$\sum_{i,l_1,l_2}\mathbb{E}\left(\left[Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right]^2\Big|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right) \le \frac{c_1^2\lambda_{\max}\mu^4}{nps^2}.$$

Note that in Assumption 1(e), we assume that the entries of the response tensor are observed independently. Thus, for all $i \in [n]$, $\delta_{i,l_1,l_2,l}$ are independent with each other. Let $X_i = \sum_{l_1,l_2}\left(Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right)$. Then,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n}X_i\right\| \ge t\right) = \mathbb{E}\left[\underbrace{\mathbb{P}\left(\left(\left\|\sum_{i=1}^{n}X_i\right\| \ge t\right)\Big|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right)}_{(*)}\right].$$

Hence, following the proof of Zhou et al, 2021 [1], we apply the Bernstein's inequality given in Theorem A.1 to (*) with $t = \gamma$ and $M = \frac{c_1^2\mu^4}{nps^2}$. We have

$$\mathbb{P}\left(\left|\sum_{i,l_1,l_2} Z_{i,l_1,l_2} - \sum_{i,l_1,l_2} \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right| \geq \gamma\right) \leq 2\exp\left\{\frac{-\gamma^2/2}{\lambda_{\max}+\gamma/3} \times \frac{pns^2}{c_1^2\mu^4}\right\},$$

where γ is a fixed positive constant:

$$\gamma = \frac{1}{2}\min\left\{\frac{\lambda_{\min}}{2}, c_2, \frac{\lambda_{\min}^2}{48\sqrt{10}\lambda_{\max}}, \frac{\lambda_{\min}^3}{48\sqrt{10}c_2\lambda_{\max}}\right\}. \tag{A.5}$$

Since $\gamma < \frac{\lambda_{\min}}{4}$, we have $\gamma < 3\lambda_{\max}$. Note that $\lambda_{\max} + \frac{\gamma}{3} > \frac{2\gamma}{3}$. Also, $\frac{\frac{-\gamma^2}{2}}{\lambda_{\max}+\frac{\gamma}{3}} > \frac{-3\gamma}{4}$. Then,

$$\frac{-\gamma^2/2}{\lambda_{\max}+\gamma/3} \times \frac{pns^2}{c_1^2\mu^4} > \frac{-3\gamma}{4} \times \frac{pns^2}{c_1^2\mu^4}.$$

By Assumption 2(d), we have that $n \geq \frac{c_5\sigma^2 s^2 \log(d)}{w_1^{*2} p}$. Therefore,

$$\frac{-\gamma^2/2}{\lambda_{\max}+\gamma/3} \times \frac{pns^2}{c_1^2\mu^4} > \frac{-3\gamma}{4} \times \frac{s^2 c_5\sigma^2 s^2 \log(d)}{c_1^2\mu^4 w_1^{*2}}.$$

Let $c' = \frac{-3\gamma s^2 c_5\sigma^2 s^2}{4c_1^2\mu^4 w_1^{*2}}$ be some negative constant, since $c_5 > 0$ and $\gamma > 0$ by the definition. Then

$$\frac{-\gamma^2/2}{\lambda_{\max}+\gamma/3} \times \frac{pns^2}{c_1^2\mu^4} > c'\log(d).$$

Hence, we have

$$1 - 2\exp\left\{\frac{-\gamma^2/2}{\lambda_{\max}+\gamma/3} \times \frac{pns^2}{c_1^2\mu^4}\right\} < 1 - 2\exp(c'\log(d)) \leq 1 - \frac{2}{d^{10}}.$$

Then, with probability at least $1-2/d^{10}$, where $d = \max\{d_1, d_2\}$, the inequality below holds:

$$\left|\sum_{i,l_1,l_2} Z_{i,l_1,l_2} - \sum_{i,l_1,l_2} \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right| \leq \gamma$$

$$\Rightarrow \left|p^{-1}\sum_{i=1}^{n} n^{-1}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2 - \sum_{i=1}^{n} n^{-1}\hat{\alpha}_{i,1}^2\right| \leq \gamma$$

$$\Rightarrow -\gamma \leq p^{-1}\sum_{i=1}^{n} n^{-1}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2 - \sum_{i=1}^{n} n^{-1}\hat{\alpha}_{i,1}^2$$

$$\Rightarrow \frac{1}{p}\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}^2}{n} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2 \geq \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}^2}{n} - \gamma > 0.$$

Therefore,

$$\left\| \left\{ \frac{1}{p} A \right\}^{-1} \right\| \leq \frac{1}{\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}^2}{n} - \gamma}. \tag{A.6}$$

Next, we find an upper bound:

$$\left\| \frac{1}{p} \left\{ B - A \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \right\} \beta_{1,3}^* \right\|.$$

Denote

$$Z_{i,l_1,l_2,l} = \left\{ \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \hat{\alpha}_{i,1}^2 \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \hat{\beta}_{1,1,l_1}^2 \hat{\beta}_{1,2,l_2}^2 \right.$$

$$\left. - \hat{\alpha}_{i,1} \alpha_{i,1}^* \beta_{1,1,l_1}^* \beta_{1,2,l_2}^* \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2} \right\} \delta_{i,l_1,l_2,l} p^{-1} n^{-1} \beta_{1,3,l}^* e_l,$$

where e_l is the d_3-column vector whose l^{th} entry is 1, others are 0 and d_3 is the dimension of $\beta_{1,3}$.
By definitions of $Z_{i,l_1,l_2,l}, A$ and B, we get that

$$\frac{1}{p} \left\{ B - A \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \right\} \beta_{1,3}^* = \sum_{i,l_1,l_2,l} Z_{i,l_1,l_2,l}.$$

Note that

$$\mathbb{E}(Z_{i,l_1,l_2} | \hat{\alpha}_{i,1} \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2}) = \left\{ \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \hat{\alpha}_{i,1}^2 \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \hat{\beta}_{1,1,l_1}^2 \hat{\beta}_{1,2,l_2}^2 \right.$$

$$\left. - \hat{\alpha}_{i,1} \alpha_{i,1}^* \beta_{1,1,l_1}^* \beta_{1,2,l_2}^* \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2} \right\} n^{-1} \beta_{1,3,l}^* e_l.$$

Then, using the fact that $\|\delta_{i,l_1,l_2,l} p^{-1} - 1\| \leq 1$, we have

$$\|Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2} | \hat{\alpha}_{i,1} \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2})\| \leq \frac{1}{np} \left| \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2} \beta_{1,3,l}^* \hat{\alpha}_{i,1} \right| \times$$

$$\left| \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1} \alpha_{i,1}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2 / n} \hat{\alpha}_{i,1} \langle \beta_{1,1}^*, \hat{\beta}_{1,1} \rangle \langle \beta_{1,2}^*, \hat{\beta}_{1,2} \rangle \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2} - \alpha_{i,1}^* \beta_{1,1,l_1}^* \beta_{1,2,l_2}^* \right|.$$

By Assumption 1, we have that $\left|\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta^*_{1,3,l}\hat{\alpha}_{i,1}\right| \leq c_1\mu^3/s^{1.5}$. This gives

$$\|Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\|$$

$$\leq \frac{c_1\mu^3}{nps^{1.5}}\left|\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n}\hat{\alpha}_{i,1}\langle\beta^*_{1,1},\hat{\beta}_{1,1}\rangle\langle\beta^*_{1,2},\hat{\beta}_{1,2}\rangle\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2} - \alpha^*_{i,1}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2}\right|$$

$$= \frac{c_1\mu^3}{nps^{1.5}}\sqrt{J_1}.$$

Therefore, by applying Proposition A.6, we have:

$$\|Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\| \leq \frac{c_1\epsilon}{np}\frac{\mu^3}{s^{1.5}}\sqrt{\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2}.$$

Also, we have

$$J_2 = \sum_{i,l_1,l_2}\mathbb{E}\left(\left[Z_{i,l_1,l_2} - \mathbb{E}(Z_{i,l_1,l_2}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right]^2\left|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right.\right)$$

$$= \frac{1}{p}\sum_{i,l_1,l_2}\frac{1}{n^2}\hat{\beta}^2_{1,1,l_1}\hat{\beta}^2_{1,2,l_2}\beta^{*2}_{1,3,l}\hat{\alpha}^2_{i,1}\left\{\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n}\hat{\alpha}_{i,1}\langle\beta^*_{1,1},\hat{\beta}_{1,1}\rangle\langle\beta^*_{1,2},\hat{\beta}_{1,2}\rangle\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right.$$

$$\left. - \alpha^*_{i,1}\beta^*_{1,1,l_1}\beta^*_{1,2,l_2}\right\}^2.$$

By Proposition A.7, we get

$$J_2 \leq \frac{\lambda_{\max}}{np}\frac{\mu^3}{s^{1.5}}\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2\epsilon^2.$$

Recall that by Assumption 1 (e) for all $i \in [n]$, $\delta_{i,l_1,l_2,l}$ are independent with each other. Let $X_i = n\sum_{l_1,l_2,l} Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})$. Then,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \geq t\right) = \mathbb{E}\left[\underbrace{\mathbb{P}\left(\left(\left\|\sum_{i=1}^n X_i\right\| \geq t\right)\left|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\right.\right)}_{(*)}\right].$$

Hence, following the proof of Zhou et al, 2021 [1], we apply the vector Bernstein's inequality given in Theorem A.2 to (*) with $\epsilon \equiv \gamma\epsilon, \sigma^2 \equiv \frac{\lambda_{\max}}{p}\frac{\mu^3}{s^{1.5}}\left\{4\frac{\lambda_{\max}^2}{\lambda_{\min}^2}+1\right\}c_1^2\epsilon^2$. We have:

$$\mathbb{P}\left(\left\|\sum_{i,l_1,l_2,l} Z_{i,l_1,l_2,l} - \sum_{i,l_1,l_2,l}\mathbb{E}(Z_{i,l_1,l_2,l}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2})\right\| \geq \gamma\epsilon\right)$$

$$\leq \exp\left\{\frac{1}{4} - \frac{\gamma^2}{8\lambda_{\max}\mu^3\left\{4\frac{\lambda_{\max}^2}{\lambda_{\min}^2}+1\right\}c_1^2/(pns^{1.5})}\right\}.$$

Note that $\sum_{i,l_1,l_2,l}\mathbb{E}(Z_{i,l_1,l_2,l}|\hat{\alpha}_{i,1}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}) = 0$. By Assumption 2(a): $p \geq c_4\{\log(d)\}^4\mu^3/\{ns^{1.5}\} > c\mu^3\log(d)/\{ns^{1.5}\gamma^2\}$ for some positive constant c. Then, the following holds with probability at least $1 - e^{\frac{1}{4}}/d^{10}$,

$$\left\|\sum_{i,l_1,l_2,l} Z_{i,l_1,l_2,l}\right\| \leq \gamma\epsilon.$$

$$\left\|\frac{1}{p}\left\{B - A\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*,\hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*,\hat{\beta}_{1,2}\rangle\right\}\beta_{1,3}^*\right\| \leq \gamma\epsilon. \tag{A.7}$$

The bound for II_1 is simplified to:

$$\|II_1\| \leq \left|\frac{w_1^*}{\hat{w}_1}\right|\left\|\left\{\frac{1}{p}A\right\}^{-1}\right\|\left\|\frac{1}{p}\left\{B - A\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*,\hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*,\hat{\beta}_{1,2}\rangle\right\}\beta_{1,3}^*\right\|.$$

Then, by combining (A.1), (A.6) and (A.7), we conclude that with probability at least $1 - e^{\frac{1}{4}}/d^{10}$,

$$\|II_1\| \leq \frac{2\gamma\epsilon}{\sum_i \hat{\alpha}_{i,1}^2/n - \gamma}.$$

Since $\sum_i \hat{\alpha}_{i,1}^2/n - \gamma = \hat{\beta}_{1,4}^T\frac{\sum_i x_i x_i^T}{n}\hat{\beta}_{1,4} - \gamma \geq \lambda_{\min} - \gamma$ and $\lambda_{\min} - \gamma > 0$ by Assumption 2(c), with probability at least $1 - e^{\frac{1}{4}}/d^{10}$, we have

$$\|II_1\| \leq \frac{2\gamma\epsilon}{\lambda_{\min} - \gamma}.$$

Note that by (A.5) $1/(\lambda_{\min} - \gamma) < 2/\lambda_{\min}$. Then, with probability at least $1 - e^{\frac{1}{4}}/d^{10}$, we get

$$\|II_1\| \leq \frac{4\gamma\epsilon}{\lambda_{\min}}.$$

□

Proposition A.9. Suppose that the conditions of Theorem 5.1 hold and let

$$III_{12} = \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \{\hat{\beta}_{1,1} - \beta_{1,1}^*\} \times_2 \hat{\beta}_{1,2}.$$

Then, with probability at least $1 - 3d^{-10}$,

$$\|III_{12}\| \leq \tilde{C}_1 \sigma \sqrt{\frac{s\log(d)}{np}} \epsilon + \tilde{C}_1 \frac{\sigma \log(d)}{np\sqrt{s}} \log\left(\sqrt{\frac{s}{p}}\right) \epsilon + \tilde{C}_2 \sigma \sqrt{\frac{s\log(d)}{np}} \epsilon^2$$

$$+ \tilde{C}_2 \frac{\sigma s \log(d)}{np} \log\left(\sqrt{\frac{s^3}{p}}\right) \epsilon^2.$$

Proof. From Assumption 2(c) and Cauchy-Schwartz inequality, we have

$$\|III_{12}\| \leq \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \hat{\beta}_{1,2} \right\| \|\hat{\beta}_{1,1} - \beta_{1,1}^*\| \leq \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \hat{\beta}_{1,2} \right\| \epsilon.$$

It suffices to bound $\left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \hat{\beta}_{1,2} \right\|$. Let $P_{\beta_{1,2}^*} = \beta_{1,2}^* \beta_{1,2}^{*T}$ - the projection onto the column space of $\beta_{1,2}^*$ and $P_{\beta_{1,2}^*}^{\perp}$ - orthogonal compliment of $P_{\beta_{1,2}^*}$. We write

$$\left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \hat{\beta}_{1,2} \right\| = \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \{P_{\beta_{1,2}^*} + P_{\beta_{1,2}^*}^{\perp}\} \hat{\beta}_{1,2} \right\|$$

$$\leq \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 P_{\beta_{1,2}^*} \hat{\beta}_{1,2} \right\| + \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 P_{\beta_{1,2}^*}^{\perp} \hat{\beta}_{1,2} \right\|. \tag{A.8}$$

Note that $\|P_{\beta_{1,2}^*} \hat{\beta}_{1,2}\| = \|\beta_{1,2}^* \beta_{1,2}^{*T} \hat{\beta}_{1,2}\|$ and $|\beta_{1,2}^{*T} \hat{\beta}_{1,2}| \leq 1$. Hence,

$$\left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 P_{\beta_{1,2}^*} \hat{\beta}_{1,2} \right\| = \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta_{1,2}^* \beta_{1,2}^{*T} \hat{\beta}_{1,2} \right\|$$

$$= |\beta_{1,2}^{*T} \hat{\beta}_{1,2}| \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta_{1,2}^* \right\| \leq \left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta_{1,2}^* \right\|.$$

Therefore, it is suffices to bound $\left\| \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta_{1,2}^* \right\|$.
We write

$$\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta_{1,2}^* = \frac{1}{pn} \sum_{i \in [n], j \in F_1, k \in F_2, l \in F_3} \alpha_{i,1}^* \delta_{i,j,k,l} \mathcal{E}_{i,j,k,l} \beta_{1,2,k}^* e_{j,l},$$

where $F_1 = \text{supp}(\beta^*_{1,1}) \cup \text{supp}(\hat{\beta}^*_{1,1})$, $F_2 = \text{supp}(\beta^*_{1,2}) \cup \text{supp}(\hat{\beta}^*_{1,2})$, $\text{supp}(v)$ refers to the set of indices in v that are nonzero and $e_{j,l} \in \mathbb{R}^{d_1 \times d_2}$ is a matrix with all zero entries except the (j,l)-th entry to 1. Let $F = F_1 \cup F_2$.
Observe that since $\|\alpha^{*2}_{i,1}\| \leq c_1^2, \|\beta_{1,2}\| = 1$, and by Corollary A.1 $\mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l}) \leq 4\sqrt{p}\sigma^2$, we have

$$\left\| \sum_{i\in[n], j\in F_1, k\in F_2, l\in F_3} \mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l})\alpha^{*2}_{i,1}\beta^{*2}_{1,2,k} e_{j,l}e^T_{j,l} \right\| \leq 4\sqrt{p}nc_1^2 s\sigma^2 \text{ and}$$

$$\left\| \sum_{i\in[n], j\in F_1, k\in F_2, l\in F_3} \mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l})\alpha^{*2}_{i,1}\beta^{*2}_{1,2,k} e^T_{j,l}e_{j,l} \right\| \leq 4\sqrt{p}nc_1^2 s\sigma^2.$$

Also, since $\beta^*_{1,2}$ is a μ-mass vector:

$$\| \|\alpha^*_{i,1}\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}\beta^*_{1,2,k}e_{j,l}\| \|_{\psi_2} \leq \|\mathcal{E}_{i,j,k,l}\|_{\psi_2} \|\alpha^*_{i,1}\beta^*_{1,2,k}e_{j,l}\| \leq c_1\sigma\frac{\mu}{\sqrt{s}},$$

where $\|\cdot\|_{\psi_2}$ is an Orlicz norm defined in chapter 2.
By Lemma A.1 with $X_i = \frac{\alpha^*_{i,1}}{p}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta^*_{1,2}$, $U^\alpha = \frac{c_1\sigma\mu}{p\sqrt{s}}$ and $\delta^2 = \frac{nc_1^2 s\sigma^2}{\sqrt{p}}$, the following bound holds with probability at least $1 - d^{-10}$,

$$\left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 \beta^*_{1,2} \right\| \leq \tilde{C}_1 \max\left\{ \sigma\sqrt{\frac{s\log(d)}{np}}, \frac{\sigma\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right) \right\} \tag{A.9}$$

for some large enough constant $\tilde{C}_1$.
Next, we bound the second term in (A.8). By Cauchy-Schwartz inequality, we have

$$\left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 P^\perp_{\beta^*_{1,2}}\hat{\beta}_{1,2} \right\| \leq \left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| \left\| (\beta^{*\perp}_{1,2})^T\hat{\beta}_{1,2} \right\|$$

$$\leq \left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| \left\| \underbrace{(\beta^{*\perp}_{1,2})^T(\beta^*_{1,2}}_{=0} - \hat{\beta}_{1,2}) \right\| \leq \left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| \left\| (\beta^{*\perp}_{1,2})^T \right\| \left\| \beta^*_{1,2} - \hat{\beta}_{1,2} \right\|.$$

Recall that $\left\|(\beta^{*\perp}_{1,2})^T\right\| = 1$ and $\left\|\beta^*_{1,2} - \hat{\beta}_{1,2}\right\| \leq \epsilon$. Then

$$\left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_2 P^\perp_{\beta^*_{1,2}}\hat{\beta}_{1,2} \right\| \leq \left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| \epsilon.$$

We write

$$\left\| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| = \sup_{u_1\in S_1, u_2\in S_2, u_3\in S_3} \left| \sum_{i=1}^n \frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 u_1 \times_2 u_2 \times_3 u_3 \right|,$$

where $S_i = \{u \in \mathbb{R}^{d_i} : \|u\| = 1, \|u\|_0 \leq s_i\}$, for $i = 1, 2, 3$ is the unit sphere in $\mathbb{R}^{d_i}$. The idea is to use a covering number algorithm. For each given subset $U_i \subseteq [d_i]$, we define the set $S_{U_i} = \{v \in \mathbb{R}^{d_i} : \|v\| = 1, \text{supp}(v) \subseteq U_i\}$. Let C_1, C_2, C_3 be $\tilde{\epsilon}$-covers of $S_{U_1}, S_{U_2}, S_{U_3}$. Next we use Lemma A.2 with $\tilde{\epsilon} = \log(3/2)/3$. We get that

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\right\| \geq \bar{t}\right) \leq \sum_{\bar{u}_j \in C_j} \mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \bar{u}_1 \times_2 \bar{u}_2 \times_3 \bar{u}_3\right| \geq \bar{t}/2\right)$$

For each fixed $\bar{u}_1, \bar{u}_2, \bar{u}_3$, we write:

$$\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \bar{u}_1 \times_2 \bar{u}_2 \times_3 \bar{u}_3 = \frac{1}{pn} \sum_{i\in[n], j\in\bar{F}_1, k\in\bar{F}_2, l\in\bar{F}_3} \alpha_{i,1}^* \delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}\bar{u}_{1,j}\bar{u}_{2,k}\bar{u}_{3,l},$$

where $\bar{F}_j = \text{supp}(\bar{u}_j)$. Since $\|\alpha_{i,1}^{*2}\| \leq c_1^2$,, and by Corollary A.1 we have that $\mathbb{E}(\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}^2) \leq 4\sqrt{p}\sigma^2$. Then, we get

$$\sum_{i\in[n], j\in\bar{F}_1, k\in\bar{F}_2, l\in\bar{F}_3} \mathbb{E}(\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}^2)\alpha_{i,1}^{*2}\bar{u}_{1,j}^2\bar{u}_{2,k}^2\bar{u}_{3,l}^2 \leq 4\sqrt{p}c_1^2 n\sigma^2$$

and

$$\|\alpha_{i,1}^*\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}\bar{u}_{1,j}\bar{u}_{2,k}\bar{u}_{3,l}\|_{\psi_2} \leq c_1\|\mathcal{E}_{i,j,k,l}\|_{\psi_2} \leq c_1\sigma,$$

Let $\bar{t} = \tilde{C}_2 \text{max}\left\{\frac{\sigma\sqrt{t}}{\sqrt{np}}, \frac{\sigma t}{np}\log\left(\sqrt{\frac{s_1 s_2 s_3}{p}}\right)\right\}$. By Lemma A.1, we have that

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \bar{u}_1 \times_2 \bar{u}_2 \times_3 \bar{u}_3\right| \geq \bar{t}\right) \leq e^{-t}.$$

Then, by properties of $\tilde{\epsilon}$-covers with $\tilde{\epsilon} = \log(3/2)/3$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\right\| \geq \bar{t}\right) \leq \sum_{\bar{u}_1\in C_1, \bar{u}_2\in C_2, \bar{u}_3\in C_3} e^{-t} \leq \left\{\frac{6}{\log(3/2)}\right\}^{s_1+s_2+s_3} 2e^{-t}.$$

Taking a union bound over $\binom{d_1}{s_1}\binom{d_2}{s_2}\binom{d_3}{s_3} \leq d^{s_1+s_2+s_3}$ choices of $U_1 \circ U_2 \circ U_3$, we get

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\right\| \geq \bar{t}\right) \leq 2e^{-t+(s_1+s_2+s_3)\log(d)}.$$

Let $t = s\log(d)$. Then, the following bound holds with probability at least $1 - 2d^{-10s}$,

$$\left\|\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\right\| \leq \bar{t} = \tilde{C}_2 \text{max}\left\{\frac{\sigma\sqrt{s\log(d)}}{\sqrt{np}}, \frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s_1 s_2 s_3}{p}}\right)\right\} \quad \text{(A.10)}$$

Combining (A.8), (A.9), (A.10), we can construct an upper bound for III_{12}. The following bound holds with probability at least $1 - 3d^{-10}$,

$$\|III_{12}\| \leq \tilde{C}_1\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_1\frac{\sigma\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right)\epsilon$$

$$+\tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon^2 + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s^3}{p}}\right)\epsilon^2.$$

□

Proposition A.10. Suppose that the conditions of Theorem 5.1 hold and let

$$III_{14} = \sum_{i=1}^{n}\frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta^*_{1,1} \times_2 \beta^*_{1,2}.$$

Then, with probability at least $1 - d^{-10}$,

$$\|III_{14}\| \leq \tilde{C}_3\max\left\{\sigma\sqrt{\frac{s\log(d)}{np}}, \frac{\sigma\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\right\},$$

for some large enough constant $\tilde{C}_3$.

Proof. We write

$$\sum_{i=1}^{n}\frac{\alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta^*_{1,1} \times_2 \beta^*_{1,2} = \frac{1}{pn}\sum_{i\in[n],j\in F_1,k\in F_2,l\in F_3}\alpha^*_{i,1}\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}\beta^*_{1,1,k}\beta^*_{1,2,k}e_l,$$

where e_l is a d_3-column vector with all zero entries except the l^{th} entry equal to 1. Let $F = F_1 \cup F_2$. Observe that, since $\|\alpha^{*2}_{i,1}\| \leq c_1^2, \|\beta_{1,2}\| = 1$, and by Corollary A.1 we have that $\mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l}) \leq 4\sqrt{p}\sigma^2$. Then, we have

$$\left\|\sum_{i\in[n],j\in F_1,k\in F_2,l\in F_3}\mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l})\alpha^{*2}_{i,1}\beta^{*2}_{1,1,j}\beta^{*2}_{1,2,k}e_le_l^T\right\| \leq 4\sqrt{p}c_1^2n\sigma^2s$$

and

$$\left\|\sum_{i\in[n],j\in F_1,k\in F_2,l\in F_3}\mathbb{E}(\delta_{i,j,k,l}\mathcal{E}^2_{i,j,k,l})\alpha^{*2}_{i,1}\beta^{*2}_{1,1,j}\beta^{*2}_{1,2,k}e_l^Te_l\right\| \leq 4\sqrt{p}c_1^2n\sigma^2s.$$

Also, since $\beta^*_{1,2}$ and $\beta^*_{1,1}$ are μ-mass vectors,

$$\|\|\alpha^*_{i,1}\delta_{i,j,k,l}\mathcal{E}_{i,j,k,l}\beta^*_{1,1,j}\beta^*_{1,2,k}e_l\|\|_{\psi_2} \leq \|\mathcal{E}_{i,j,k,l}\|_{\psi_2}\|\alpha^*_{i,1}\beta^*_{1,1,j}\beta^*_{1,2,k}e_l\| \leq \frac{c_1\mu^2\sigma}{s}.$$

Set $X_i = \frac{\alpha^*_{i,1}}{p}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta^*_{1,1} \times_2 \beta^*_{1,2}$, $U^\alpha = \frac{c_1\sigma\mu^2}{ps}$ and $\delta^2 = \frac{nc_1^2 s\sigma^2}{\sqrt{p}}$. By Lemma A.1, with probability at least $1 - d^{-10}$,

$$\|III_{14}\| \leq \tilde{C}_3 \max\left\{\sigma\sqrt{\frac{s\log(d)}{np}}, \frac{\sigma\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\right\},$$

for some large enough constant $\tilde{C}_3$.

□

Proposition A.11. Suppose that the conditions of Theorem 5.1 hold and let

$$III_{11} = \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} - \alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}.$$

Then, with probability at least $1 - 2d^{-10}$,

$$\|III_{11}\| \leq \tilde{C}_3\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_3\frac{\sigma\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\epsilon$$

$$+\tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon^2 + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s_1 s_2 s_3}{p}}\right)\epsilon^2.$$

Proof.

$$\|III_{11}\| = \left\|\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} - \alpha^*_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}\right\|$$

$$= \left\|\sum_{i=1}^{n} \frac{\hat{\beta}^T_{1,4}x_i - \beta^{*T}_{1,4}x_i}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}\right\|.$$

Then, by Cauchy-Schwartz inequality, we get

$$\|III_{11}\| \leq \left\|\sum_{i=1}^{n} \frac{1}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}x_i^T\right\| \left\|\hat{\beta}_{1,4} - \beta^*_{1,4}\right\|.$$

We write

$$\sum_{i=1}^{n} \frac{1}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}x_i^T$$

$$= \sum_{i=1}^{n} \frac{1}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \{P_{\beta^*_{1,1}} + P^{\perp}_{\beta^*_{1,1}}\}\hat{\beta}_{1,1} \times_2 \{P_{\beta^*_{1,2}} + P^{\perp}_{\beta^*_{1,2}}\}\hat{\beta}_{1,2}x_i^T$$

$$= \sum_{i=1}^{n} \frac{1}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 P_{\beta^*_{1,1}}\hat{\beta}_{1,1} \times_2 P_{\beta^*_{1,2}}\hat{\beta}_{1,2}x_i^T + \hat{D}, \tag{A.11}$$

with

$$\hat{D} = \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 P_{\beta^*_{1,1}} \hat{\beta}_{1,1} \times_2 P^{\perp}_{\beta^*_{1,2}} \hat{\beta}_{1,2} x_i^T$$

$$+ \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 P^{\perp}_{\beta^*_{1,1}} \hat{\beta}_{1,1} \times_2 P_{\beta^*_{1,2}} \hat{\beta}_{1,2} x_i^T + \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 P^{\perp}_{\beta^*_{1,1}} \hat{\beta}_{1,1} \times_2 P^{\perp}_{\beta^*_{1,2}} \hat{\beta}_{1,2} x_i^T.$$

Since $\|\beta^{*T}_{1,1} \hat{\beta}_{1,1}\| \leq 1$, it is sufficient to find an upper bound of

$$\sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta^*_{1,1} \times_2 \beta^*_{1,2} x_i^T.$$

Observe that since $\|\alpha^{*2}_{i,1}\| \leq c_1^2, \|\beta_{1,2}\| = 1$, and by Corollary A.1 we have that $\mathbb{E}(\delta_{i,j,k,l} \mathcal{E}^2_{i,j,k,l}) \leq 4\sqrt{p}\sigma^2$. Then, we get

$$\left\| \sum_{i\in[n], j\in F_1, k\in F_2, l\in F_3} \mathbb{E}(\delta_{i,j,k,l} \mathcal{E}^2_{i,j,k,l}) \beta^{*2}_{1,1,l} \beta^{*2}_{1,2,k} e_l x_i^T x_i e_l^T \right\| \leq 4\sqrt{p}\sigma^2 \left\| \sum_{i\in[n], l\in F_3} e_l x_i^T x_i e_l^T \right\|.$$

Then

$$\left\| \sum_{i\in[n], j\in F_1, k\in F_2, l\in F_3} \mathbb{E}(\delta_{i,j,k,l} \mathcal{E}^2_{i,j,k,l}) \beta^{*2}_{1,1,l} \beta^{*2}_{1,2,k} e_l x_i^T x_i e_l^T \right\| \leq 4\sqrt{p} c_1^2 n \sigma^2,$$

$$\left\| \sum_{i\in[n], j\in F_1, k\in F_2, l\in F_3} \mathbb{E}(\delta_{i,j,k,l} \mathcal{E}^2_{i,j,k,l}) \beta^{*2}_{1,1,j} \beta^{*2}_{1,2,k} x_i e_l^T e_l x_i^T \right\| \leq p\sigma^2 s \left\| \sum_{i\in[n]} x_i x_i^T \right\| \leq 4\sqrt{p} c_1^2 n \sigma^2 s,$$

Also, since $\beta^*_{1,2}$ and $\beta^*_{1,1}$ are μ-mass vectors:

$$\| \| \delta_{i,j,k,l} \mathcal{E}_{i,j,k,l} \beta^*_{1,1,j} \beta^*_{1,2,k} e_l x_i^T \| \|_{\psi_2} \leq \frac{c_1 \mu^2 \sigma}{s}.$$

By Lemma A.1, in the same way as in Proposition A.10, with probability at least $1 - d^{-10}$,

$$\left\| \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta^*_{1,1} \times_2 \beta^*_{1,2} x_i^T \right\| \leq \tilde{C}_3 \left\{ \sigma \sqrt{\frac{s \log(d)}{np}}, \frac{\sigma \log(d)}{nps} \log\left(\frac{1}{\sqrt{p}}\right) \right\},$$

for some large enough constant $\tilde{C}_3$. Next we prove the upper bound of $\hat{D}$ in (A.11). We have

$$J_3 = \left\| \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 P_{\beta^*_{1,1}} \hat{\beta}_{1,1} \times_2 P^{\perp}_{\beta^*_{1,2}} \hat{\beta}_{1,2} x_i^T \right\| \leq c_1 \left\| \sum_{i=1}^{n} \frac{1}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\| \|(\beta^{*\perp}_{1,2})^T \hat{\beta}_{1,2}\|.$$

Recall that $\|(\beta_{1,2}^{*\perp})^T\hat{\beta}_{1,2}\| \leq \epsilon$. Since $\alpha_{i,1}^*$ is a scalar and $|\alpha_{i,1}^*| \leq c_1$, an upper bound for $\left\|\sum_{i=1}^{n}\frac{1}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\right\|$ is similar to the bound from (A.10). Then, with probability at least $1-2/d^{10}$

$$J_3 \leq \tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s_1s_2s_3}{p}}\right)\epsilon.$$

The other terms in the $\hat{D}$ are bounded similarly. Therefore, with probability at least $1-d^{-10}$ we have

$$\|\hat{D}\| \leq \tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s_1s_2s_3}{p}}\right)\epsilon.$$

Hence, we conclude that with probability at least $1-2/d^{10}$, we have

$$\|III_{11}\| \leq \tilde{C}_3\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_3\frac{\sigma\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\epsilon$$

$$+\tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon^2 + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s_1s_2s_3}{p}}\right)\epsilon^2.$$

□

Proposition A.12. Suppose that the conditions of Theorem 5.1 hold and let

$$III_1 = \frac{1}{\hat{w}_1}A^{-1}\left\{\sum_{i=1}^{n}\frac{\hat{\alpha}_{i,1}}{n}\Pi_{\Omega_i}(\mathcal{E}_{iF})\times_1\hat{\beta}_{1,1}\times_2\hat{\beta}_{1,2}\right\}.$$

Then, with probability at least $1-d^{-10}$, we have

$$\|III_1\| \leq \frac{2\tilde{C}\sigma}{w_1^*\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}} + 6\gamma'\epsilon,$$

for some large enough $\tilde{C} > 0$.

Proof.

$$\|III_1\| \leq \left|\frac{1}{\hat{w}_1}\right|\left\|\left\{\frac{1}{p}A\right\}^{-1}\right\|\left\|\sum_{i=1}^{n}\frac{\hat{\alpha}_{i,1}}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\times_1\hat{\beta}_{1,1}\times_2\hat{\beta}_{1,2}\right\|.$$

In (A.6) we proved that

$$\left\|\left\{\frac{1}{p}A\right\}^{-1}\right\| \leq \frac{1}{\sum_{i=1}^{n}\frac{\hat{\alpha}_{i,1}^2}{n}-\gamma}.$$

Since from (A.1) $\hat{w}_1 > \frac{w_1^*}{2}$, we have that

$$\|III_1\| \leq \frac{2}{w_1^*\{\lambda_{\min} - \gamma\}} \left\| \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2} \right\|. \tag{A.12}$$

Next, we need to bound the term $\left\| \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2} \right\|$. We write

$$\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}$$

$$= \sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} - \alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2} + \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}. \tag{A.13}$$

Similarly, we can write

$$\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}$$

$$= \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \{\hat{\beta}_{1,1} - \beta_{1,1}^*\} \times_2 \hat{\beta}_{1,2} + \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \hat{\beta}_{1,2}. \tag{A.14}$$

Also,

$$\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \hat{\beta}_{1,2}$$

$$= \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \{\hat{\beta}_{1,2} - \beta_{1,2}^*\} + \sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \beta_{1,2}^*. \tag{A.15}$$

Hence, combining equalities A.13 - A.15, we have

$$\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}$$

$$= \underbrace{\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1} - \alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}}_{III_{11}} + \underbrace{\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \{\hat{\beta}_{1,1} - \beta_{1,1}^*\} \times_2 \hat{\beta}_{1,2}}_{III_{12}}$$

$$+ \underbrace{\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \{\hat{\beta}_{1,2} - \beta_{1,2}^*\}}_{III_{13}} + \underbrace{\sum_{i=1}^{n} \frac{\alpha_{i,1}^*}{pn} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^* \times_2 \beta_{1,2}^*}_{III_{14}}.$$

From Proposition A.11, with probability at least $1 - 2d^{-10}$, $\|III_{11}\| \leq \tilde{C}_3 \sigma \sqrt{\frac{\text{slog}(d)}{np}} \epsilon + \tilde{C}_3 \frac{\sigma \text{slog}(d)}{nps} \log\left(\frac{1}{\sqrt{p}}\right) \epsilon + \tilde{C}_2 \sigma \sqrt{\frac{\text{slog}(d)}{np}} \epsilon^2 + \tilde{C}_2 \frac{\sigma \text{slog}(d)}{np} \log\left(\sqrt{\frac{s_1 s_2 s_3}{p}}\right) \epsilon^2$.

From Proposition A.9, with probability at least $1-3d^{-10}$, $\|III_{12}\| \leq \tilde{C}_1\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_1\frac{\sigma\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right)\epsilon + \tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon^2 + \tilde{C}_2\frac{\sigma s\log(d)}{np}\log\left(\sqrt{\frac{s^3}{p}}\right)\epsilon^2$.
Also, from Proposition A.10, with probability at least $1-d^{-10}$, for some large enough constant $\tilde{C}_3$, we have $\|III_{14}\| \leq \tilde{C}_3\max\left\{\sigma\sqrt{\frac{s\log(d)}{np}}, \frac{\sigma\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\right\}$.
Next, we consider III_{13}. Note that $\sum_{i=1}^{n}\frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^*$ is different from $\sum_{i=1}^{n}\frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF})\times_2\beta_{1,2}^*$ only by vectors β^*. Since $\beta_{1,1}^*$ and $\beta_{1,2}^*$ have the same properties, the bound for III_{13} can be derived in the similar way as in (A.9). Hence, we have that

$$\|III_{13}\| \leq \left\|\sum_{i=1}^{n}\frac{\alpha_{i,1}^*}{pn}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \beta_{1,1}^*\right\|\epsilon \leq \tilde{C}_1\sigma\sqrt{\frac{s\log(d)}{np}}\epsilon + \tilde{C}_1\frac{\sigma\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right)\epsilon. \tag{A.16}$$

Therefore, using Propositions A.11, A.13, A.14 and (A.16), with probability at least $1-d^{-10}$, we can bound $\|III_1\|$ as

$$\begin{aligned}\|III_1\| \leq \frac{2\{2\tilde{C}_1+\tilde{C}_3\}\sigma\epsilon}{w_1^*\{\lambda_{\min}-\gamma\}}&\left(\sqrt{\frac{s\log(d)}{np}} + \frac{\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right) + \sqrt{\frac{s\log(d)}{np}}\right.\\ &\left.+\frac{\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\right) + \frac{4\tilde{C}_2\sigma\epsilon^2}{w_1^*\{\lambda_{\min}-\gamma\}}\left(\sqrt{\frac{s\log(d)}{np}} + \frac{s\log(d)}{np}\log\left(\sqrt{\frac{s^3}{p}}\right)\right)\\ &+\frac{2\tilde{C}_3\sigma}{w_1^*\{\lambda_{\min}-\gamma\}}\sqrt{\frac{s\log(d)}{np}}.\end{aligned} \tag{A.17}$$

From (A.5), we have $\gamma < \lambda_{\min}/2$ and $1/\{\lambda_{\min}-\gamma\} < 2/\lambda_{\min}$. By Assumption 2(d), we have

$$\frac{4\{2\tilde{C}_1+\tilde{C}_3\}\sigma\epsilon}{\lambda_{\min}w_1^*}\sqrt{\frac{s\log(d)}{np}} \leq \gamma'\epsilon,$$

for some positive constant γ' that will be determined later. Similarly, we have the following bounds:

$$\frac{4\{2\tilde{C}_1+\tilde{C}_3\}\sigma\epsilon}{\lambda_{\min}w_1^*}\max\left\{\frac{\log(d)}{np\sqrt{s}}\log\left(\sqrt{\frac{s}{p}}\right), \frac{\log(d)}{nps}\log\left(\frac{1}{\sqrt{p}}\right)\right\} \leq \gamma'\epsilon,$$

$$\frac{8\tilde{C}_2\sigma\epsilon^2}{w_1^*\lambda_{\min}}\max\left\{\sqrt{\frac{s\log(d)}{np}}, \frac{s\log(d)}{np}\log\left(\sqrt{\frac{s^3}{p}}\right)\right\} \leq \gamma'\epsilon.$$

Using the bound above, (A.17) and (A.5), with probability at least $1-d^{-10}$, we have

$$\|III_1\| \leq \frac{2\tilde{C}\sigma}{w_1^*\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}} + 6\gamma'\epsilon,$$

for some large enough $\tilde{C} > 0$.

□

Proposition A.13. Suppose that the conditions of Theorem 5.1 hold. Let $A_1 = \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T$ and $A_2 = \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T$. Then, A_1 and A_2 are positive definite matrix.

Proof. Let $u \in \mathbb{R}^q, u \neq 0$. Consider

$$u^T A_1 u = u^T \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T u = \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 (u^T x_i)(x_i^T u)$$

$$= \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 (x_i^T u)^T (x_i^T u) = \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 |x_i^T u|^2.$$

If $u^T A_1 u = 0$, we have that

$$\sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 |x_i^T u|^2 = 0.$$

Note that by the model (1) and Assumptions on B^*, $\hat{\beta}_{1,1} \neq 0, \hat{\beta}_{1,2} \neq 0$ and $\hat{\beta}_{1,3} \neq 0$. Moreover, there is exists at least one i, such that $\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F \neq 0$. Thus, $\sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 > 0$. Let $K = \sum_{i=1}^{n} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 > 0$ and $\mathbb{P}(A) = \sum_{i \in A} \frac{\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2}{K}$, where A is a Borel set of $\mathbb{R}$. We have

$$\sum_{i=1}^{n} \frac{\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2}{K} \left|x_i^T u\right|^2 = \sum_{i=1}^{n} \left|x_i^T u\right|^2 \mathbb{P}\left[U = \left|x_i^T u\right|\right] = \mathbb{E}\left[U^2\right],$$

where U is a random variable taking the values $\left|x_1^T u\right|, \left|x_2^T u\right|, \ldots, \left|x_n^T u\right|$ with $\mathbb{P}\left[U = \left|x_i^T u\right|\right] = \frac{\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2}{K}$, $i = 1, 2, \ldots, n$. Then, $u^T A_1 u = 0$ if and only if $\mathbb{E}[U^2] = 0$. This holds if and only if $U^2 = 0$ almost surely with respect to the probability measure $\mathbb{P}$. Then, $\left|x_i^T u\right|^2 = 0$ for all $i \in [n]$. Therefore, we have

$$\sum_{i=1}^{n} |x_i^T u|^2 = 0 \Rightarrow u^T \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T u = 0.$$

By assumption 1(a), we get

$$\lambda_{\min} u^T u \leq u^T \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T u \leq \lambda_{\max} u^T u.$$

Since $\lambda_{\min} > 0$, we have

$$u^T u = 0.$$

Therefore $u = 0$ and then $u^T A_1 u > 0$ for any $u \in \mathbb{R}^q, u \neq 0$, this proves that A_1 is a positive definite matrix.

Next, consider

$$u^T A_2 u = u^T \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T u = \sum_{i=1}^{n} \frac{1}{n} (u^T x_i)(x_i^T u) = \sum_{i=1}^{n} \frac{1}{n} (x_i^T u)^T (x_i^T u) = \sum_{i=1}^{n} \frac{1}{n} |x_i^T u|^2.$$

If $u^T A_2 u = 0$, we have that

$$u^T \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T u = 0.$$

By assumption 1(a), we get

$$\lambda_{\min} u^T u \leq u^T \sum_{i=1}^{n} \frac{1}{n} x_i x_i^T u \leq \lambda_{\max} u^T u.$$

Since $\lambda_{\min} > 0$, we have

$$u^T u = 0.$$

Therefore $u = 0$ and then $u^T A_2 u > 0$ for any $u \in \mathbb{R}^q, u \neq 0$, this proves that A_2 is a positive definite matrix. □

Proposition A.14. Suppose that the conditions of Theorem 5.1 hold and let

$$I_2 = \left\| \left\{ \frac{1}{np} \sum_{i=1}^{n} \left\| \Pi_{\Omega_i}(\hat{A}_1) \right\|_F^2 x_i x_i^T \right\}^{-1} \right\|.$$

Then, with probability at least $1 - 2q/d^{10}$, we have $I_2 \leq \frac{8}{\lambda_{\min} w_1^{*2}}$.

Proof. At first, we show that

$$\left\| \frac{1}{n} \left\{ \sum_{i=1}^{n} \left(\frac{1}{p} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 - \|\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3}\|_F^2 \right) \right\} x_i x_i^T \right\| \leq \gamma,$$

where γ is the same constant as defined as (A.5). Assume that for all $i \in [n]$, $j \in [d_1], k \in [d_2], l \in [d_3]$, $\delta_{i,j,k,l}$ is independent with $\hat{\beta}_{1,1,j}, \hat{\beta}_{1,2,k}, \hat{\beta}_{1,3,l}$. Further, let

$$Z_i = \frac{1}{n} \left\{ \frac{1}{p} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 - \|\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3}\|_F^2 \right\} x_i x_i^T.$$

Since $\|\hat{\beta}_{1,j}\| = 1, j \in [3]$ and $\|\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3}\|_F^2 = \|\hat{\beta}_{1,1}\|_F^2 \|\hat{\beta}_{1,2}\|_F^2 \|\hat{\beta}_{1,3}\|_F^2 = 1$, we have

$$Z_i = \frac{1}{n} \left\{ \frac{1}{p} \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 - 1 \right\} x_i x_i^T.$$

Then, since $\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 \leq 1$ and $\|x_i\| \leq c_1$, we have

$$\|Z_i\| \leq \frac{1}{n}\left|\frac{1}{p}\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 - 1\right| \|x_i x_i^T\| \leq c_1^2 \frac{1}{n}\left|\frac{1}{p} - 1\right| \leq \frac{c_1^2}{p}.$$

In addition, we have

$$\begin{aligned}
&\left\|\sum_{i=1}^n \mathbb{E}(Z_i^2|\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3})\right\| \\
&= \left\|\sum_{i=1}^n \frac{1}{n^2}\mathbb{E}\left(\left\{\frac{1}{p}\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 - 1\right\}^2 |\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3}\right) x_i x_i^T x_i x_i^T\right\| \\
&= \left\|\sum_{i=1}^n \frac{1}{n^2}\mathbb{E}\left(\left\{\frac{1}{p}\sum_{j,k,l}\delta_{i,j,k,l}\hat{\beta}_{1,1,j}^2\hat{\beta}_{1,2,k}^2\hat{\beta}_{1,3,l}^2 - 1\right\}^2 |\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3}\right) x_i x_i^T x_i x_i^T\right\|
\end{aligned}$$

Since $\mathbb{E}(\delta_{i,j,k,l}) = p$, this gives

$$\left\|\sum_{i=1}^n \mathbb{E}(Z_i^2|\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3})\right\| \leq \left\|\sum_{i=1}^n \sum_{j,k,l} \frac{1}{n^2 p}\hat{\beta}_{1,1,j}^4\hat{\beta}_{1,2,k}^4\hat{\beta}_{1,3,l}^4 x_i x_i^T x_i x_i^T\right\|,$$

and then, using the facts that $|\hat{\beta}_{k,j,l} \leq \frac{\mu}{\sqrt{s}}$, $\|x_i\| \leq c_1$, $\frac{1}{n}\sum_i \|x_i x_i^T\|_2 \leq c_2$, we get

$$\left\|\sum_{i=1}^n \mathbb{E}(Z_i^2|\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3})\right\| \leq \frac{1}{np}\sum_{j,k,l}\hat{\beta}_{1,1,j}^4\hat{\beta}_{1,2,k}^4\hat{\beta}_{1,3,l}^4 \left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^T x_i x_i^T\right\| \leq \frac{c_1^2 c_2 \mu^3}{nps^{1.5}}.$$

Note that in Assumption 1 (e) we assume that the entries of the response tensor are observed independently. Let $X_i = \frac{1}{p}\|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T - x_i x_i^T$. Then,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right) = \mathbb{E}\left[\underbrace{\mathbb{P}\left(\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right)\middle|\hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3}\right)}_{(*)}\right].$$

Hence, following the proof of Zhou et al, 2021 [1], we apply the matrix Bernstein's inequality given in Theorem A.3 to (*) with $\sigma^2 = \frac{c_1^2 c_2 \mu^3}{nps^{1.5}}$, $B = \frac{c_1^2}{p}, t = \gamma$ and $d_1 = d_2 = q$. We have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{np}\sum_i \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T - \sum \frac{1}{n} x_i x_i^T\right\| \leq \gamma\right) \\
&\geq 1 - 2q\exp\left\{\frac{-\gamma^2/2}{c_1^2 c_2 \mu^3/\{nps^{1.5}\} + c_1^2\gamma/\{3p\}}\right\}.
\end{aligned}$$

By Assumption 2(a), $p \geq c_4\mu^3\{\log(d)\}/(ns^{1.5}) \geq c\mu^3\{\log(d)\}/(ns^{1.5}\gamma^2)$ for some constant c. Then, with probability at least $1 - 2q/d^{10}$, we have

$$\left\| \frac{1}{np} \sum_i \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T - \sum \frac{1}{n} x_i x_i^T \right\| \leq \gamma. \tag{A.18}$$

Let $A_1 = \sum_i \|\Pi_{\Omega_i}(\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3})\|_F^2 x_i x_i^T$ and $A_2 = \sum_i \frac{1}{n} x_i x_i^T$. By the Proposition A.13, A_1 and A_2 are positive definite matrix. Hence, A_1 and A_2 are invertible.

To apply Proposition A.3, we first check that $1 - \|A_1 - A_2\|\|A_2^{-1}\| > 0$.

$$1 - \|A_1 - A_2\|\|A_2^{-1}\| = 1 - \gamma \left\| \left\{ 1/n \sum_{i=1}^n x_i x_i^T \right\}^{-1} \right\| = 1 - \frac{\gamma}{\lambda_{\min}}.$$

Recall that by (A.5), we have $\gamma < \lambda_{\min}/2$. Then

$$1 - \|A_1 - A_2\|\|A_2^{-1}\| > 1 - \frac{1}{2} = \frac{1}{2} > 0.$$

Using Proposition A.3, we get

$$I_2 = \left\| \left\{ \frac{1}{np} \sum_{i=1}^n \left\| \Pi_{\Omega_i}(\hat{A}_1) \right\|_F^2 x_i x_i^T \right\}^{-1} \right\| \leq \frac{1}{\hat{w}_1^2} \frac{\left\| \left\{ 1/n \sum_{i=1}^n x_i x_i^T \right\}^{-1} \right\|}{1 - \gamma \left\| \left\{ 1/n \sum_{i=1}^n x_i x_i^T \right\}^{-1} \right\|},$$

and then using (A.18), (A.5) and the fact that $|\hat{w}_1 - w_1^*| < w_1^*/2$, we have

$$I_2 \leq \frac{4}{\lambda_{\min} w_1^{*2}\{1 - \gamma/\lambda_{\min}\}} \leq \frac{8}{\lambda_{\min} w_1^{*2}}.$$

□

Proposition A.15. Suppose that the conditions of Theorem 5.1 hold and let

$$II_2 = \left\| \frac{1}{np} \sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A_1^* - \hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1) \right\rangle x_i x_i^T \beta_{1,4}^* \right\|.$$

Then, with probability at least $1 - e^{\frac{1}{4}/d^{10}}$

$$II_2 \leq \{6c_2 + \gamma\} w_1^{*2} \epsilon.$$

Proof. Let $Z_{i,l_1,l_2,l} = \frac{1}{np} \delta_{i,l_1,l_2,l} (A_1^* - \hat{A}_1)_{i,l_1,l_2,l} (\hat{A}_1)_{i,l_1,l_2,l} x_i x_i^T \beta_{1,4}^*$. Then

$$\frac{1}{np} \sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A_1^* - \hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1) \right\rangle x_i x_i^T \beta_{1,4}^* = \sum_{i,l_1,l_2,l} Z_{i,l_1,l_2,l}.$$

Assume that for all $i \in [n], l_1 \in [d_1], l_2 \in [d_2], l \in [d_3]$, $\delta_{i,l_1,l_2,l}$ is independent with $\hat{A}_1$. Note that, since $\mathbb{E}(\delta_{i,l_1,l_2,l}) = p$, we get

$$\begin{aligned}\|Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{A}_1)\| &= \|Z_{i,l_1,l_2,l} - \frac{1}{n}(A_1^* - \hat{A}_1)_{i,l_1,l_2,l}(\hat{A}_1)_{i,l_1,l_2,l}x_i x_i^T \beta_{1,4}^*\| \\ &\leq |\delta_{i,l_1,l_2,l} - p| \left\| \frac{1}{np}(A_1^* - \hat{A}_1)_{i,l_1,l_2,l}(\hat{A}_1)_{i,l_1,l_2,l}x_i x_i^T \beta_{1,4}^* \right\|.\end{aligned}$$

Then,

$$\|Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{A}_1)\| \leq \frac{1}{pn} \left|(A_1^* - \hat{A}_1)_{i,l_1,l_2,l}\right| \left|(\hat{A}_1)_{i,l_1,l_2,l}\right| \left\|x_i x_i^T \beta_{1,4}^*\right\|. \quad \text{(A.19)}$$

By Assumption 1, we have $\left\|x_i x_i^T \beta_{1,4}^*\right\| \leq c_1^2$. Also, since $|\hat{w}_1| \leq \frac{3}{2}w_1^*$ and $\hat{\beta}$'s are μ-mass vectors, we have $\left|(\hat{A}_1)_{i,l_1,l_2,l}\right| \leq \frac{3}{2}w_1^*\mu^3/s^{1.5}$. Furthermore, by triangle inequality, we have

$$\begin{aligned}&\left|(A_1^* - \hat{A}_1)_{i,l_1,l_2,l}\right| \leq \left|w_1^*\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*\beta_{1,3,l}^* - \hat{w}_1\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*\beta_{1,3,l}^*\right| \\ &+\left|\hat{w}_1\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*\beta_{1,3,l}^* - \hat{w}_1\hat{\beta}_{1,1,l_1}\beta_{1,2,l_2}^*\beta_{1,3,l}^*\right| + \left|\hat{w}_1\hat{\beta}_{1,1,l_1}\beta_{1,2,l_2}^*\beta_{1,3,l}^* - \hat{w}_1\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta_{1,3,l}^*\right| \\ &\left|\hat{w}_1\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta_{1,3,l}^* - \hat{w}_1\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\hat{\beta}_{1,3,l}\right| \leq \frac{6\mu^2 w_1^*\epsilon}{s}. \quad \text{(A.20)}\end{aligned}$$

Therefore,

$$\|Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{A}_1)\| \leq \frac{9w_1^{*2}\mu^5 c_1^2\epsilon}{pns^{2.5}}.$$

Also, from (A.19), (A.20) and Assumption 1(a), we have

$$\begin{aligned}\sum_{i,l_1,l_2,l} \mathbb{E}(\|Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{A}_1)\|^2|\hat{A}_1) &= \sum_{i,l_1,l_2,l} \mathbb{E}(\|Z_{i,l_1,l_2,l} \\ &-\frac{1}{n}(A_1^* - \hat{A}_1)_{i,l_1,l_2,l}(\hat{A}_1)_{i,l_1,l_2,l}x_i x_i^T \beta_{1,4}^*\|^2) \\ &\leq \frac{1}{pn}\sum_i \frac{1}{n}\beta_{1,4}^{*T}x_i x_i^T x_i x_i^T \beta_{1,4}^* \frac{9w_1^{*2}\mu^6}{4s^3}\left\{6w_1^*\epsilon\frac{\mu^2}{s}\right\}^2,\end{aligned}$$

and then

$$\sum_{i,l_1,l_2,l} \mathbb{E}(\|Z_{i,l_1,l_2,l} - \mathbb{E}(Z_{i,l_1,l_2,l}|\hat{A}_1)\|^2|\hat{A}_1) \leq \frac{81c_1^2c_2w_1^{*4}\mu^3\epsilon^2}{pns^{1.5}}.$$

Note that in Assumption 1(e) we assume that the entries of the response tensor are observed independently. Thus, for all $i \in [n]$, $\delta_{i,l_1,l_2,l}$ are independent with each other. Let $X_i = \sum_{l_1,l_2,l} Z_{i,l_1,l_2,l} - \frac{1}{n}\left\langle A_1^* - \hat{A}_1, \hat{A}_1 \right\rangle x_i x_i^T \beta_{1,4}^*$. Then

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right) = \mathbb{E}\left[\underbrace{\mathbb{P}\left(\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right)\Big|\hat{A}_1\right)}_{(*)}\right].$$

Following the proof of Zhou et al, 2021 [1], we apply the vector Bernstein's inequality given in Theorem A.2 to (*) with $\sigma^2 \equiv \frac{81c_1^2c_2w_1^{*4}\mu^3\epsilon^2}{ps^{1.5}}, \epsilon \equiv \gamma w_1^{*2}\epsilon$. Hence, we have

$$\mathbb{P}\left(\left\|\sum_{i,l_1,l_2,l} Z_{i,l_1,l_2,l} - \frac{1}{n}\sum_{i=1}^{n}\left\langle A_1^* - \hat{A}_1, \hat{A}_1\right\rangle x_i x_i^T \beta_{1,4}^*\right\| \leq \gamma w_1^{*2}\epsilon\right)$$

$$\geq 1 - \exp\left(\frac{1}{4} - \frac{\gamma^2}{\frac{8\times 81c_1^2c_2\mu^3}{pns^{1.5}}}\right).$$

By Assumption 2(a) $p \geq c\mu^3\{\log(d)\}/(ns^{1.5}\gamma^2)$ for some constant c. Therefore, with probability at least $1 - e^{\frac{1}{4}/d^{10}}$,

$$II_2 \leq \frac{1}{n}\sum_i \langle A_1^* - \hat{A}_1, \hat{A}_1\rangle x_i x_i^T \beta_{1,4}^* + \gamma w_1^{*2}\epsilon \leq \frac{1}{n}\sum_i \left\|A_1^* - \hat{A}_1\right\|_F \left\|\hat{A}_1\right\|_F \|x_i x_i^T\| + \gamma w_1^{*2}\epsilon.$$

Note that $\|\hat{A}_1\|_F = |\hat{w}_1| \leq 3w_1^*/2$. We next bound $\|A_1^* - \hat{A}_1\|_F$. By triangle inequality, we have

$$\begin{aligned}
\|A_1^* - \hat{A}_1\|_F &= \left\|\hat{w}_1\hat{\beta}_{1,1}\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3} - w_1^*\beta_{1,1}^*\circ\beta_{1,2}^*\circ\beta_{1,3}^*\right\|_F \\
&\leq \underbrace{\left\|\hat{w}_1\hat{\beta}_{1,1}\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3} - w_1^*\hat{\beta}_{1,1}\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3}\right\|_F}_{II_{21}} \\
&+ \underbrace{\left\|w_1^*\hat{\beta}_{1,1}\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3} - w_1^*\beta_{1,1}^*\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3}\right\|_F}_{II_{22}} \\
&+ \underbrace{\left\|w_1^*\beta_{1,1}\circ\hat{\beta}_{1,2}\circ\hat{\beta}_{1,3} - w_1^*\beta_{1,1}\circ\beta_{1,2}\circ\hat{\beta}_{1,3}\right\|_F}_{II_{23}} \\
&+ \underbrace{\left\|w_1^*\beta_{1,1}\circ\beta_{1,2}\circ\hat{\beta}_{1,3} - w_1^*\beta_{1,1}\circ\beta_{1,2}\circ\beta_{1,3}\right\|_F}_{II_{24}}.
\end{aligned}$$

Note that $II_{21} = |\hat{w}_1 - w_1^*| < w_1^*\epsilon, II_{22} \leq w_1^*\|\hat{\beta}_{1,1} - \beta_{1,1}^*\| \leq w_1^*\epsilon, II_{23} \leq w_1^*\epsilon$ and $II_{24} \leq w_1^*\epsilon$. Therefore, we have $\|A_1^* - \hat{A}_1\|_F \leq 4w_1^*\epsilon$. By Assumption 1, we have $\frac{1}{n}\sum_{i=1}^{n}\|x_i x_i^T\| \leq c_2$. Hence,

$$II_2 \leq 4w_1^*\epsilon\frac{3}{2}w_1^*c_2 + \gamma w_1^{*2}\epsilon = \{6c_2 + \gamma\}w_1^{*2}\epsilon.$$

□

Proposition A.16. Suppose that the conditions of Theorem 5.1 hold and let

$$III_2 = \left\|\frac{1}{np}\sum_i \langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1)\rangle x_i\right\|.$$

Then, with probability at least $1 - 2q/d^{10s}$

$$III_2 \leq \frac{3\tilde{C}_2\sigma w_1^*}{2}\sqrt{\frac{qs\log(d)}{np}}.$$

Proof. Note that, since Π_{Ω_i} is an indicator tensor, $\left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1) \right\rangle = \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_1 \right\rangle$. We have

$$III_2 = \left\| \frac{1}{np} \sum_i \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_1 \rangle x_i \right\| = \left\| \hat{w}_1 \frac{1}{np} \sum_{i=1}^n \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3} \rangle x_i \right\|. \quad \text{(A.21)}$$

The j^{th} entry of the vector $\frac{1}{np}\sum_i \langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1)\rangle x_i$ for each $j \in [q]$ can be written as

$$\frac{1}{np} \sum_i \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_1 \rangle x_{i,j} = \frac{c_1}{np} \sum_i \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_1 \rangle x_{i,j}/c_1.$$

Our goal is to find the upper bound of

$$\left\| \frac{1}{np} \sum_{i=1}^n \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3} \rangle x_{i,j} \right\|.$$

Since $|x_{i,j}/c_1| \leq 1$, it is sufficient to bound

$$\left\| \frac{c_1}{np} \sum_{i=1}^n \langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3} \rangle \right\| \leq \left\| \frac{c_1}{np} \sum_{i=1}^n \Pi_{\Omega_i}(\mathcal{E}_i) \right\|. \quad \text{(A.22)}$$

By following the similar method as the upper bound of $\leq \left\| \frac{1}{np} \sum_{i=1}^n \alpha_{i,1}^* \Pi_{\Omega_i}(\mathcal{E}_{iF}) \right\|$ in (A.10), we obtain for each $j \in [q]$, with probability at least $1 - 2d^{-10s}$, that

$$\left\| \frac{c_1}{np} \sum_{i=1}^n \Pi_{\Omega_i}(\mathcal{E}_i) \right\| \leq \tilde{C}_2\sigma\sqrt{\frac{s\log(d)}{np}}, \quad \text{(A.23)}$$

for $\tilde{C}_2 > 0$. Therefore, by combining (A.21), (A.22), (A.23), along with the fact that $|\hat{w}_1| \leq \frac{3}{2}w_1^*$, we conclude that with probability at least $1 - 2q/d^{10s}$

$$III_2 \leq \frac{3\tilde{C}_2\sigma w_1^*}{2}\sqrt{\frac{qs\log(d)}{np}}.$$

□

Proposition A.17. Suppose that the conditions of Theorem 5.2 hold and let

$$II_1 = \frac{1}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}_{i,k}^2/n} \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,k} \sum_{k' \neq k} \Bigg\{ \alpha_{i,k'}^* w_{k',1}^* \langle \bar{\beta}_{k',1}^*, \hat{\beta}_{k,1} \rangle \langle \bar{\beta}_{k',2}^*, \hat{\beta}_{k,2} \rangle \bar{\beta}_{k',3}^*$$

$$-\hat{\alpha}_{i,k'}\hat{w}^*_{k',1}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3}\Bigg\}.$$

Then,

$$\|II_1\| \leq \frac{c_1^2 r w^*_{\max}\{8\epsilon^2 + 8\xi\epsilon + 5\xi^2\epsilon\}}{\lambda_{\min} w^*_{\min}}.$$

Proof. Note that $\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle = \langle\bar{\hat{\beta}}^*_{k',1} - \bar{\beta}^*_{k',1} + \bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2} - \bar{\beta}^*_{k',2} + \bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle$. Also, from Assumptions 3(c, d) and triangle inequality $\left|\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\right| = \left|\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1} - \beta^*_{k,1} + \beta^*_{k,1}\rangle\right| \leq \left|\langle\bar{\beta}^*_{k',1},\beta^*_{k,1}\rangle\right| + \left|\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1} - \beta^*_{k,1}\rangle\right| \leq \xi + \epsilon$.
Hence, we have

$$\begin{aligned}
&\left\|\sum_{k'\neq k}\left\{\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\beta}^*_{k',3} - \alpha^*_{i,k'}w^*_{k',1}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3}\right\}\right\| \\
&\qquad \leq \left\|\sum_{k'\neq k}\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\{\bar{\beta}^*_{k',3} - \bar{\hat{\beta}}^*_{k',3}\}\right\| \\
&\qquad\quad + \left\|\sum_{k'\neq k}\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2} - \bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3}\right\| \\
&\qquad\quad + \left\|\sum_{k'\neq k}\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\langle\bar{\hat{\beta}}^*_{k',1} - \bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\bar{\hat{\beta}}^*_{k',3}\right\| \\
&\qquad + \left\|\sum_{k'\neq k}\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\hat{\beta}}^*_{k',2} - \bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\langle\bar{\hat{\beta}}^*_{k',1} - \bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\bar{\hat{\beta}}^*_{k',3}\right\| \\
&\leq c_1 r w^*_{\max}\{\xi+\epsilon\}^2\epsilon + 2c_1 r w^*_{\max}\{\xi+\epsilon\}\epsilon + c_1 r w^*_{\max}\epsilon^2 \leq 4c_1 r w^*_{\max}\epsilon^2 + 4c_1 r w^*_{\max}\xi\epsilon.
\end{aligned}$$

Besides,

$$\begin{aligned}
&\left\|\sum_{k'\neq k}\left\{\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3} - \hat{\alpha}_{i,k'}\hat{w}_{k',1}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3}\right\}\right\| \\
&\leq r|\alpha^*_{i,k'}w^*_{k'} - \hat{\alpha}_{i,k'}\hat{w}_{k'}|\xi^2 \leq |\hat{\alpha}_{i,k'}\hat{w}_{k'} - \hat{w}_{k'}\alpha^*_{i,k'} + \hat{w}_{k'}\alpha^*_{i,k'} - \alpha^*_{i,k'}w^*_{k'}|r\xi^2 \\
&\leq \left(|\hat{w}_{k'}||\hat{\alpha}_{i,k'} - \alpha^*_{i,k'}| + |\alpha^*_{i,k'}||\hat{w}_{k'} - w^*_{k'}|\right) r\xi^2 \\
&\leq \left(c_1\epsilon w^*_{\max} + c_1\epsilon w^*_{\max}\right) r\xi^2 \leq \frac{5c_1 r\epsilon w^*_{\max}\xi}{2}.
\end{aligned}$$

Then, together with the triangle inequality, we get

$$\|II_1\| \leq \frac{2c_1^2}{\lambda_{\min}w^*_{\min}}\left[4c_1 r w^*_{\max}\epsilon^2 + 4c_1 r w^*_{\max}\xi\epsilon + \frac{5c_1 r\epsilon w^*_{\max}\xi}{2}\right].$$

Then,

$$\|II_1\| \leq \frac{c_1^2 r w^*_{\max}\{8\epsilon^2 + 8\xi\epsilon + 5\xi^2\epsilon\}}{\lambda_{\min} w^*_{\min}}.$$

□

Proposition A.18. Suppose that the conditions of Theorem 5.2 hold and let

$$IV_1 = \sum_{k' \neq k} \frac{w^*_{k'}}{\hat{w}_k} A^{-1} \left\{ F_{k'} - A \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1}\rangle \langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2}\rangle \right\} \bar{\beta}^*_{k',3}$$

$$- \sum_{k' \neq k} \frac{\hat{w}^*_{k'}}{\hat{w}_k} A^{-1} \left\{ G_{k'} - A \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\hat{\alpha}_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \langle \bar{\hat{\beta}}^*_{k',1}, \hat{\beta}_{k,1}\rangle \langle \bar{\hat{\beta}}^*_{k',2}, \hat{\beta}_{k,2}\rangle \right\} \bar{\hat{\beta}}^*_{k',3},$$

where $A, F_{k'}$ and $G_{k'}$ are diagonal matrices with diagonal entry,

$$A_{ll} = \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{\beta}^2_{k,1,l_1} \hat{\beta}^2_{k,2,l_2}$$

$$F_{k'll} = \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,k}\alpha^*_{i,k'} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \bar{\beta}^*_{k',1,l_1} \bar{\beta}^*_{k',2,l_2}$$

$$G_{k'll} = \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,k}\hat{\alpha}_{i,k'} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \bar{\hat{\beta}}^*_{k',1,l_1} \bar{\hat{\beta}}^*_{k',2,l_2}. \tag{A.24}$$

Then, with probability at least $1 - 2/d^{10}$, $\|IV_1\| \leq \frac{4\gamma\epsilon}{\lambda_{\min}}$.

Proof. Denote

$$Z_{i,k',l_1,l_2,l} = \frac{1}{pn} \frac{w^*_{k'}}{\hat{w}_k} \left\{ \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \hat{\alpha}^2_{i,k} \langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1}\rangle \langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2}\rangle \hat{\beta}^2_{k,1,l_1} \hat{\beta}^2_{k,2,l_2} \right.$$

$$\left. - \hat{\alpha}_{i,k}\alpha^*_{i,k'} \bar{\beta}_{k',1,l_1} \bar{\beta}_{k',2,l_2} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \right\} \delta_{i,l_1,l_2,l} \bar{\beta}^*_{k',3,l} e_l,$$

where e_l is the column vector whose l^{th} entry is 1 and others are 0. Then, we have

$$\frac{1}{p} \sum_{k' \neq k} \frac{w^*_k}{\hat{w}_k} \left\{ F_{k'} - A \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1}\rangle \langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2}\rangle \right\} \bar{\beta}^*_{k',3} = \sum_{i,k',l_1,l_2,l} Z_{i,k',l_1,l_2,l}.$$

Then, following the similar steps from Proposition A.8, we have

$$\|Z_{i,k',l_1,l_2,l} - \mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})\| \leq \frac{2w^*_{\max}}{w^*_{\min}}\frac{1}{pn}\left|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\bar{\beta}^*_{k',3,l}\hat{\alpha}_{i,k}\right| J_{1,k}^{\frac{1}{2}},$$

where

$$J_{1,k} = \left| \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\hat{\alpha}_{i,k}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2} - \alpha^*_{i,k'}\bar{\beta}_{k',1,l_1}\bar{\beta}_{k',2,l_2}\right|^2. \tag{A.25}$$

Applying Proposition A.6 and following steps in (A.7), we get

$$\|Z_{i,k',l_1,l_2,l} - \mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})\| \leq \frac{2w^*_{\max}}{w^*_{\min}pn}\frac{\mu^3}{s^{1.5}}c_1\epsilon\sqrt{\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2}.$$

In addition, we have

$$\sum_{i,k',l_1,l_2,l}\mathbb{E}(\|Z_{i,k',l_1,l_2,l} - \mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})\|^2|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k}) =$$

$$= \left\{\frac{1}{p}-1\right\}\sum_{i,k',l_1,l_2,l}\frac{4w^{*2}_{\max}}{w^{*2}_{\max}n^2}\hat{\beta}^2_{k,1,l_1}\hat{\beta}^2_{k,2,l_2}\bar{\beta}^{*2}_{k',3,l}\hat{\alpha}^2_{i,k}J_{1,k},$$

where $J_{1,k}$ is given in (A.25). Then

$$\sum_{i,k',l_1,l_2,l}\mathbb{E}(\|Z_{i,k',l_1,l_2,l} - \mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})\|^2|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})$$

$$\leq \frac{4w^{*2}_{\max}}{w^{*2}_{\max}n^2}\sum_i\frac{\hat{\alpha}^2_{i,k}\mu^4}{ns^2}\sum_{k',l_1,l_2}\sum_l\bar{\beta}^{*2}_{k',3,l}J_{1,k} \leq \frac{4rw^{*2}_{\max}\lambda_{\max}\mu^3}{w^{*2}_{\max}pns^{1.5}}\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2\epsilon^2.$$

Recall that in Assumption 1(e) we assume that the entries of the response tensor are observed independently. Thus, for all $i \in [n]$, $\delta_{i,l_1,l_2,l}$ are independent with each other. Let $X_i = \sum_{k',l_1,l_2,l}(Z_{i,k',l_1,l_2,l} - \mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k}))$. Then

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right) = \mathbb{E}\left[\underbrace{\mathbb{P}\left(\left(\left\|\sum_{i=1}^n X_i\right\| \leq t\right)\Big|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k}\right)}_{(*)}\right].$$

Then, following the proof of Zhou et al, 2021 [1], we apply the vector Bernstein's inequality given in Theorem A.2 to (*) with $\mu \equiv \frac{2w^*_{\max}}{w^*_{\min}pn}\frac{\mu^3}{s^{1.5}}c_1\epsilon\sqrt{\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2}$ and

$\sigma^2 \equiv \frac{4rw^{*2}_{\max}\lambda_{\max}\mu^3}{w^{*2}_{\max}pns^{1.5}}\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2\epsilon^2$. Note that $\sum_{i,k',l_1,l_2,l}\mathbb{E}(Z_{i,k',l_1,l_2,l}|\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\hat{\alpha}_{i,k})=0$. Therefore,

$$\mathbb{P}\left(\left\|\sum_{i,k',l_1,l_2,l}Z_{i,k',l_1,l_2,l}\right\|\leq\gamma\epsilon\right)\geq 1-\max\left\{\frac{1}{4}-\frac{n\gamma^2}{8\frac{4rw^{*2}_{\max}\lambda_{\max}\mu^3\left\{4\frac{\lambda^2_{\max}}{\lambda^2_{\min}}+1\right\}c_1^2}{w^{*2}_{\max}pns^{1.5}}}\right\}.$$

By Assumption 3(a), $p\geq c_7 r\mu^3 w^{*2}_{\max}\log(d)/(ns^{1.5}w^{*2}_{\min})\geq cr\mu^3 w^{*2}_{\max}\log(d)/(ns^{1.5}w^{*2}_{\min}\gamma^2)$ for some positive constant c. Then, with probability at least $1-e^{1/4}/d^{10}$,

$$\left\|\frac{1}{p}\sum_{k'\neq k}\frac{w^*_k}{\hat{w}_k}\left\{F_{k'}-A\frac{\sum_{i=1}^n\hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n\hat{\alpha}^2_{i,k}/n}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\right\}\bar{\beta}^*_{k',3}\right\|\leq\gamma\epsilon.$$

Similarly, with probability at least $1-e^{1/4}/d^{10}$, we have,

$$\left\|\frac{1}{p}\sum_{k'\neq k}\frac{\hat{w}_{k'}}{\hat{w}_k}\left\{G_{k'}-A\frac{\sum_{i=1}^n\hat{\alpha}_{i,k}\hat{\alpha}'_{i,k}/n}{\sum_{i=1}^n\hat{\alpha}^2_{i,k}/n}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\right\}\bar{\hat{\beta}}^*_{k',3}\right\|\leq\gamma\epsilon.$$

Note that

$$\|IV_1\|\leq\left\|\frac{1}{p}A\right\|^{-1}\left(\left\|\frac{1}{p}\sum_{k'\neq k}\frac{w^*_k}{\hat{w}_k}\left\{F_{k'}-A\frac{\sum_{i=1}^n\hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n\hat{\alpha}^2_{i,k}/n}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\right\}\bar{\beta}^*_{k',3}\right\|\right.$$

$$\left.+\left\|\frac{1}{p}\sum_{k'\neq k}\frac{\hat{w}_{k'}}{\hat{w}_k}\left\{G_{k'}-A\frac{\sum_{i=1}^n\hat{\alpha}_{i,k}\hat{\alpha}'_{i,k}/n}{\sum_{i=1}^n\hat{\alpha}^2_{i,k}/n}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\right\}\bar{\hat{\beta}}^*_{k',3}\right\|\right)\leq\left\|\frac{1}{p}A\right\|^{-1}2\gamma\epsilon.$$

As we have shown in (A.6), if $p\geq c_5\mu^4\log(d)/\{ns^2\}\geq c\mu^4\log(d)/\{ns^2\gamma^2\}$ for some positive constant c, then, with probability at least $1-2/d^{10}$, each entry of the diagonal matrix A has the lower bound $|1/pA_{ll}|\geq\sum_{i=1}^n\hat{\alpha}^2_{i,1}/n-\gamma\geq\lambda_{\min}-\gamma$.
Therefore, IV_1 can be bounded as

$$\|IV_1\|\leq\frac{2\gamma\epsilon}{\lambda_{\min}-\gamma}\leq\frac{4\gamma\epsilon}{\lambda_{\min}}.$$

□

Proposition A.19. Suppose that Assumptions 2(d) holds and let g be a d-column random vector, such that

$$g \sim \mathrm{N}(0, I_d).$$

Then, with probability at least $1 - \mathcal{O}(d^{-12})$, $\|g\| \lesssim \sqrt{\log d}$.

Proof. By Markov's inequality, for a monotonically increasing nonnegative function h, we have

$$\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq \frac{\mathbb{E}h\left(\|g\|\right)}{h\left(\sqrt{\log d}\right)}.$$

Let $h(x) = e^{c_0 x^2}, 0 < c_0 < \frac{1}{2}$. Then, $h(x)$ is a monotonically increasing nonnegative function for $x > 0$. Therefore,

$$\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq \frac{\mathbb{E}e^{c_0\|g\|^2}}{e^{c_0 \log d}} = \frac{\mathbb{E}e^{c_0\|g\|^2}}{d^{c_0}} = \frac{M_{\|g\|^2}(c_0)}{d^{c_0}},$$

where $M_{\|g\|^2}(t)$ is a moment generating function of a random variable $\|g\|^2$. Note that, since $g \sim \mathrm{N}(0, I_d)$, $\|g\|^2$ follows chi-square distribution with d degrees of freedom. Hence, $M_{\|g\|^2}(c_0) = \frac{1}{(1-2c_0)^{\frac{d}{2}}}$ and

$$\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq d^{-c_0}(1-2c_0)^{\frac{-d}{2}} = d^{-c_0} e^{\frac{-d}{2}\log(1-2c_0)}.$$

Then, we have

$$d^{12}\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq d^{12-c_0} e^{\frac{-d}{2}\log(1-2c_0)}.$$

Recall that by Assumption 2(d), $\log(d) \leq \frac{n w_1^{*2} p}{c_5 \sigma^2 s^2}$. Let $C_5 = \exp\left(\frac{n w_1^{*2} p}{c_5 \sigma^2 s^2}\right)$. Thus, $d \leq C_5$. Consider $f(x) = x^{12-c_0} e^{\frac{-x}{2}\log(1-2c_0)}, 1 < x < C_5, 0 < c_0 < \frac{1}{2}$. Then, $f'(x) = -\frac{1}{2} x^{11-c_0} e^{\frac{-x}{2}\log(1-2c_0)} \left(x\log(1-2c_0) + 2c_0 - 24\right)$. Note that, for $1 < x < C_5, 0 < c_0 < \frac{1}{2}$, $-\frac{1}{2} x^{11-c_0} e^{\frac{-x}{2}\log(1-2c_0)} < 0$. Also, we have

$$0 < 1-2c_0 < 1 \Rightarrow \log(1-2c_0) < 0 \Rightarrow x\log(1-2c_0) < 0 \Rightarrow x\log(1-2c_0) + 2c_0 - 24 < 0.$$

Therefore, for $1 < x < C_5, 0 < c_0 < \frac{1}{2}$, $f'(x) > 0$. Thus, $f(x)$ is a monotonically increasing function on $1 < x < C_5$, which gives $f(x) < f(C_5)$ on $1 < x < C_5$. Hence, we get

$$d^{12}\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq C_5^{-c_0} e^{\frac{-C_5}{2}\log(1-2c_0)}.$$

Therefore,

$$\mathbb{P}\left(\|g\| > \sqrt{\log d}\right) \leq \mathcal{O}(d^{-12}),$$

and with probability at least $1 - \mathcal{O}(d^{-12})$, $\|g\| \lesssim \sqrt{\log d}$.

□

Lemma A.5. (Lemma 8 from Zhou et al, 2021 [1].)

Suppose Assumption 2(b): $\tau_{s_j} \geq s_j, \tau_{s_j} = \int_{|}$ and $\tau_{f_j} \geq f_j$ - holds. Then,

$$\|\hat{\beta}^f_{1,3} - \beta^*_{1,3}\| \leq \left\| \tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\| - \beta^*_{1,3} \right\|,$$

The equality holds if and only if $\hat{\beta}^f_{1,3} = \tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\|$.

Lemma A.6. (Lemma 12 from Yuan and Zhang, 2013 [16].)

Consider a sparse vector x with $\text{supp}(x) = F_x$ and $F_x = d_0$. Let $F_y = \text{supp}(y, s)$. If $\|x\| = \|y\| = 1$, then

$$|\mathit{Truncate}(y, F_y)^T x| \geq |y^T x| - \sqrt{\frac{d_0}{s}} \min \left[1 - (y^T x)^2, \left(1 + \sqrt{\frac{d_0}{s}} \right) \{1 - (y^T x)^2\} \right].$$

Lemma A.7. (Lemma S.6.2 from Sun et al, 2017 [17].)

For any tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and an index set $F = F_1 \circ F_2 \circ F_3$ with $F_i \subseteq \{1, \ldots, \}$, if $\mathcal{T} = \sum_{i \in [R]} w_i a_i \circ b_i \circ c_i$, then

$$\mathcal{T}_F = \sum_{i \in [R]} w_i \mathit{Truncate}(a_i, F_1) \circ \mathit{Truncate}(b_i, F_2) \circ \mathit{Truncate}(c_i, F_3).$$

Lemma A.8. (Lemma D.5 from Cai et al, 2021 [18].)

Let $\{X_{i,j}\}_{1 \leq i \leq r, 1 \leq j \leq L}$ be a sequence of i.i.d. standard Gaussian random variables. Consider some quantities $k \geq 1, \Delta > 0$ and $0 < \delta < \frac{1}{2}$. There exists some universal constant $C > 0$ such that if

$$L \geq C r^{2k^2} (k\sqrt{r} + \delta) \exp(\Delta^2) \log(\frac{1}{\delta}),$$

then with probability at least $1 - \delta$, there exists some $1 \leq j_0 \leq L$ such that $X_{1,j_0} > k \max_{1 < i \leq r} |X_{i,j_0}| + \Delta$.

Lemma A.9. (Lemma D.4 from Cai et al, 2021 [18].)

Suppose that $p \gtrsim d^{-2} \log^3 d$ and that $\mu \log^2 d \lesssim d$. Then for any fixed vector $w \in \mathbb{R}^d$, with probability $1 - \mathcal{O}(d^{-10})$, one has

$$\|(p^{-1}T - T^*) \times_3 w\| \lesssim \|w\|_\infty \sqrt{\frac{\mu r \log d}{dp}} \lambda^*_{\max} + \|w\|_\infty \frac{\sigma \log^{5/2} d}{p} + \|w\|_2 \sigma \sqrt{\frac{d \log d}{p}}.$$

The result also holds if we replace $\times_3$ with $\times_1$ or $\times_2$.

Lemma A.10. (Corollary D.3 from Cai et al, 2021 [18].)

With probability $1 - \mathcal{O}(d^{-10})$, one has

$$\|p^{-1}\Pi_\Omega(T^*) - T^*\| \lesssim \frac{\sqrt{\mu r}\lambda^*_{\max}\log^3 d}{d^{3/2}p} + \frac{\mu\sqrt{r}\lambda^*_{\max}\log^{5/2} d}{d\sqrt{p}} \text{ and}$$

$$\|\Pi_\Omega(E)\| \lesssim \sigma(\log^{7/2} d + \sqrt{dp}\log^{5/2} d).$$

Lemma A.11. (Based on Lemma 5.7 from Cai et al, 2021 [18].)

Suppose that the conditions of Theorem 5.3 hold. Then, with probability at least $1 - \mathcal{O}(d^{-10})$, we have

$$\|U_1 U_1^T - U_1^* U_1^{*T}\| \leq \mathcal{E}_{\text{se}} \ll \frac{1}{\sqrt{\log d}}.$$

Appendix B Proof of the main results

B.1 Proof of Theorem 5.1

The proof is for the case of $m = 2$ to simplify the calculations. However, it shows the ideas and methods used for the proof which could be extended to the case where $m > 2$.

The true model for rank $r = 1$ reduces to

$$Y_i = \omega_1^*(\beta_{1,4}^{*T})\beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^* + \mathcal{E}_i, i = 1, \ldots, n$$

The plan for the proof is to bound estimators from each step of the algorithm. Then, to combine them to bound the estimator from the i-th iteration.

Error bound of the estimator from Step 1

In the first step we obtain the unconstrained estimator $\tilde{\beta}_{1,3}$, as an example. We can use $\tilde{\beta}_{1,1}$ or $\tilde{\beta}_{1,2}$ too.

The closed form solution of the optimization problem (5) for $\tilde{\beta}_{1,3}$ was derived in Section 4.1:

$$\tilde{\beta}_{1,3,l} = \frac{\sum_{i=1}^{n}(\hat{\alpha}_{i,1})^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{R}_{i,1,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}}{\sum_{i=1}^{n}(\hat{\alpha}_{i,1})^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{w}_1(\hat{\beta}_{1,1,l_1})^2(\hat{\beta}_{1,2,l_2})^2},$$

where $\hat{R}_i = Y_i/\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,1} = \hat{\beta}_{1,4}^T x_i$.

For the above solution $\tilde{\beta}_{1,1}$ and $\tilde{\beta}_{1,2}$ are μ-mass unit vectors, since we assume that $\tilde{\beta}_{1,1}$ and $\tilde{\beta}_{1,2}$ were estimated in a previous loop.

Denote $F_1 = \text{supp}(\beta_{1,1}^*) \cup \text{supp}(\hat{\beta}_{1,1}^*)$, $F_2 = \text{supp}(\beta_{1,2}^*) \cup \text{supp}(\hat{\beta}_{1,2}^*)$, $F_3 = \text{supp}(\beta_{1,3}^*) \cup \text{supp}(\hat{\beta}_{1,3}^*)$, where $\text{supp}(v)$ refers to the set of indices in v that are nonzero.
Let $F = F_1 \cup F_2 \cup F_3$. Then, we consider the following estimator, that is equivalent to the estimator above.

$$\tilde{\beta}_{1,3,l}^* = \frac{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}(\hat{R}_{iF})_{,1,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}}{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{w}_1(\hat{\beta}_{1,1,l_1})^2(\hat{\beta}_{1,2,l_2})^2},$$

where R_{iF} denotes the restricted version of the tensor R_i on the three modes indexed by F_1, F_2 and F_3.

We note that due to the sparsity restriction and the scaling-invariant truncation operation, replacing $\tilde{\beta}_{1,3,l}$ by $\tilde{\beta}_{1,3,l}^*$ does not affect the iteration of $\tilde{\beta}_{1,3,l}$ Sun et al, 2017 [17]. Therefore, we assume that $\tilde{\beta}_{1,3}$ has been replaced by $\tilde{\beta}_{1,3}^*$. Since $Y_i = \alpha_{i,1}^* w_1^* \beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^* + \mathcal{E}_i$, we have

$$\tilde{\beta}_{1,3,l} = \frac{w_1^* \sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}\alpha_{i,1}^* \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*\beta_{1,3,l_3}^*}{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{w}_1(\hat{\beta}_{1,1,l_1})^2(\hat{\beta}_{1,2,l_2})^2}$$
$$+\frac{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}(\mathcal{E}_{iF})_{l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}}{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{w}_1(\hat{\beta}_{1,1,l_1})^2(\hat{\beta}_{1,2,l_2})^2}. \tag{B.1}$$

It can be written in a vector form as:

$$\tilde{\beta}_{1,3} = \frac{w_1^* \sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^{n} \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle\beta_{1,3}^*$$
$$+\frac{w_1^*}{\hat{w}_1}A^{-1}\left\{B - A\frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle\right\}\beta_{1,3}^*$$
$$+\frac{1}{\hat{w}_1}A^{-1}\left\{\sum_{i=1}^{n} \frac{\hat{\alpha}_{i,1}}{n}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}\right\},$$

where A and B are diagonal matrices with diagonal entry, defined in (A.4) as

$$A_{ll} = \sum_{i=1}^{n} \hat{\alpha}_{i,1}^2/n \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}^2\hat{\beta}_{1,2,l_2}^2$$
$$B_{ll} = \sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha_{i,1}^*/n \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*.$$

To prove that this vector form is equal to (B.1), let us consider an element from the vector $\tilde{\beta}_{1,3}$.

The third term is straightforward to the second term of (B.1) using the definition of mode-n multiplication. Further, let $C = \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^{n} \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle$. We have,

$$\left(\frac{w_1^*}{\hat{w}_1}C\beta_{1,3}^* + \frac{w_1^*}{\hat{w}_1}A^{-1}\left\{B - AC\right\}\beta_{1,3}^*\right)_l = \left(\frac{w_1^*}{\hat{w}_1}A^{-1}B\beta_{1,3}^*\right)_l$$
$$= \frac{w_1^* \sum_{i=1}^{n} \hat{\alpha}_{i,1}\alpha_{i,1}^*/n \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{\beta}_{1,1,l_1}\hat{\beta}_{1,2,l_2}\beta_{1,1,l_1}^*\beta_{1,2,l_2}^*\beta_{1,3,l}^*}{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,1}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l}\hat{w}_1(\hat{\beta}_{1,1,l_1})^2(\hat{\beta}_{1,2,l_2})^2}.$$

We want to bound the distance between $\tilde{\beta}_{1,3}$ and $\frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*$.

We have

$$\tilde{\beta}_{1,3} - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^* = I_1 + II_1 + III_1, \tag{B.2}$$

where

$$I_1 = \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\left\{\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle - 1\right\}\beta_{1,3}^*$$

$$II_1 = \frac{w_1^*}{\hat{w}_1}A^{-1}\left\{B - A\frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\langle\beta_{1,1}^*, \hat{\beta}_{1,1}\rangle\langle\beta_{1,2}^*, \hat{\beta}_{1,2}\rangle\right\}\beta_{1,3}^*$$

$$III_1 = \frac{1}{\hat{w}_1}A^{-1}\left\{\sum_{i=1}^n \frac{\hat{\alpha}_{i,1}}{n}\Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2}\right\}.$$

By Proposition A.5, we get that $\|I_1\| \leq \frac{2\lambda_{\max}}{\lambda_{\min}}\epsilon^2$. By Proposition A.8, with probability at least $1 - e^{\frac{1}{4}}/d^{10}$, we have $\|II_1\| \leq \frac{4\gamma\epsilon}{\lambda_{\min}}$. By Proposition A.12, with probability at least $1 - d^{-10}$, $\|III_1\| \leq \frac{2\tilde{C}\sigma}{w_1^*\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}} + 6\gamma'\epsilon$.

Hence, combining the bounds from Propositions A.5, A.8 and A.12, with probability at least $1 - 1/d^9$, we have,

$$\left\|\tilde{\beta}_{1,3} - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\| \leq \left\{\frac{2\lambda_{\max}}{\lambda_{\min}}\epsilon + \frac{4\gamma}{\lambda_{\min}} + 6\gamma'\right\}\epsilon + \frac{2\tilde{C}\sigma}{w_1^*\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}}. \tag{B.3}$$

Finally, we bound the distance between the normalized $\tilde{\beta}_{1,3}$ and the true parameter $\beta_{1,3}^*$. Since $\|\beta_{1,3}^*\| = 1$, we have:

$$\left|\|\tilde{\beta}_{1,3}\| - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right| = \left|\left\|\tilde{\beta}_{1,3}\frac{\|\tilde{\beta}_{1,3}\|}{\|\tilde{\beta}_{1,3}\|}\right\| - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\|\beta_{1,3}^*\|\right|$$

Recall that $|\|x\| - \|y\|| \leq \|x - y\|$. Therefore, this gives

$$\left|\|\tilde{\beta}_{1,3}\| - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right| \leq \left\|\tilde{\beta}_{1,3} - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\|. \tag{B.4}$$

Note that

$$\left|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|\left\|\frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|}-\beta_{1,3}^*\right\|-\left\|\tilde{\beta}_{1,3}-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\|$$

$$\leq\left\|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|}-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*-\tilde{\beta}_{1,3}+\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\|$$

$$=\left\|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|}-\tilde{\beta}_{1,3}\right\|=\left\|\tilde{\beta}_{1,3}\right\|\left|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\frac{1}{\|\tilde{\beta}_{1,3}\|}-1\right|$$

$$=\left|\|\tilde{\beta}_{1,3}\|-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|.$$

Therefore,

$$\left|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|\left\|\frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|}-\beta_{1,3}^*\right\|\leq\left|\|\tilde{\beta}_{1,3}\|-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|$$

$$+\left\|\tilde{\beta}_{1,3}-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\|.$$

By combining with (B.4), we get

$$\left|\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\right|\left\|\frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|}-\beta_{1,3}^*\right\|\leq 2\left\|\tilde{\beta}_{1,3}-\frac{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}\beta_{1,3}^*\right\|. \tag{B.5}$$

Recall that $\hat{w}_1 - w_1^* \leq \frac{1}{2}w_1^*$. Then, $\frac{\hat{w}_1}{w_1^*}\leq\frac{3}{2}$. Using (A.2), we have

$$\left|\frac{\hat{w}_1\sum_{i=1}^n \hat{\alpha}_{i,1}^2/n}{w_1^*\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha_{i,1}^*/n}\right|\leq\frac{3\lambda_{\max}}{2\lambda_{\min}}.$$

Therefore,

$$\left\| \frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|} - \beta^*_{1,3} \right\| \leq \frac{3\lambda_{\max}}{\lambda_{\min}} \left\| \tilde{\beta}_{1,3} - \frac{w_1^* \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \beta^*_{1,3} \right\|,$$

and then, from (B.3), we conclude that with probability at least $1 - 1/d^9$,

$$\left\| \frac{\tilde{\beta}_{1,3}}{\|\tilde{\beta}_{1,3}\|} - \beta^*_{1,3} \right\| \leq \frac{6\lambda_{\max}}{\lambda_{\min}} \left\{ \frac{\lambda_{\max}}{\lambda_{\min}} + \frac{2\gamma}{\lambda_{\min}} + 3\gamma' \right\} \epsilon + \frac{6\lambda_{\max}\tilde{C}\sigma}{w_1^*\lambda^2_{\min}} \sqrt{\frac{s\log(d)}{np}}. \tag{B.6}$$

Error bound of the estimator from Step 2

After obtaining the normalized vector $\tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\|$ from Step 1, we apply the *Truncatefuse* operator to $\tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\|$ to obtain the sparse and fused operator.

First, we apply the fusion operator to $\tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\|$. We have

$$\hat{\beta}^f_{1,3} = \arg\min_{\beta} \left\| \beta - \tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\| \right\| \text{ such that } \|D\beta\|_0 \leq \tau_{f_3},$$

where τ_{f_3} is the fusion parameter used in Algorithm 1, and f_3 is the true fusion parameter.

Then, by Lemma A.5, we have that, $\|\hat{\beta}^f_{1,3} - \beta^*_{1,3}\| \leq \left\| \tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\| - \beta^*_{1,3} \right\|$. In other words, if the true parameter has a fusion structure, then adding the fusion step is guaranteed to reduce the estimation error.

Next, we apply the *Truncate* operator to $\hat{\beta}^f_{1,3}$. By Lemma A.6, we have

$$\left| \mathit{Truncate}\left(\hat{\beta}^f_{1,3}, \tau_{s_3}\right)^T \beta^*_{1,3} \right| \geq \left| \hat{\beta}^{fT}_{1,3}\beta^*_{1,3} \right| - \sqrt{\frac{s_3}{f_3}} \left(1 + \sqrt{\frac{s_3}{f_3}}\right) \left\{ 1 - \left(\hat{\beta}^{fT}_{1,3}\beta^*_{1,3}\right)^2 \right\},$$

where the right-hand-side is an increasing function in terms of $|\hat{\beta}^{fT}_{1,3}\beta^*_{1,3}|$.

Let $\hat{\beta}_{1,3} = \mathit{Truncate}\left(\hat{\beta}^f_{1,3}, \tau_{s_3}\right) / \left\| \mathit{Truncate}\left(\hat{\beta}^f_{1,3}, \tau_{s_3}\right) \right\|$.

Note that $\left\| \mathit{Truncate}\left(\hat{\beta}^f_{1,3}, \tau_{s_3}\right) \right\| \leq 1$ due to the facts that $\|\hat{\beta}^f_{1,3}\| = 1$ and *Truncate* operator sets some entries in $\hat{\beta}^f_{1,3}$ to 0. Therefore,

$$\|\bar{\beta}_{1,3} - \beta^*_{1,3}\| \leq \sqrt{2}\sqrt{1 - \left(\hat{\beta}^T_{1,3}\beta^*_{1,3}\right)^2} \leq \sqrt{2}\sqrt{1 - \left\{ \mathit{Truncate}\left(\hat{\beta}^f_{1,3}, \tau_{s_3}\right)^T \beta^*_{1,3} \right\}^2}$$

$$\leq \sqrt{2} \left\{ 1 + \sqrt{\frac{s_3}{f_3}} \left(1 + \sqrt{\frac{s_3}{f_3}}\right) \right\}^{1/2} \sqrt{2}\sqrt{1 - \left(\hat{\beta}^{fT}_{1,3}\beta^*_{1,3}\right)^2},$$

and then

$$\|\bar{\beta}_{1,3} - \beta^*_{1,3}\| \leq \sqrt{10}\|\bar{\beta}^f_{1,3} - \beta^*_{1,3}\| \leq \sqrt{10} \left\| \tilde{\beta}_{1,3}/\|\tilde{\beta}_{1,3}\| - \beta^*_{1,3} \right\|. \tag{B.7}$$

Combining (B.6) and (B.7), with probability at least $1 - 1/d^9$, we have,

$$\|\bar{\beta}_{1,3} - \beta^*_{1,3}\| \leq \frac{6\sqrt{10}\lambda_{\max}}{\lambda_{\min}} \left\{ \frac{\lambda_{\max}}{\lambda_{\min}} + \frac{2\gamma}{\lambda_{\min}} + 3\gamma' \right\} \epsilon + \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w^*_1 \lambda^2_{\min}} \sqrt{\frac{s\log(d)}{np}}. \quad \text{(B.8)}$$

Finally, we need to prove that, if the true parameter $\beta^*_{1,3}$ is a μ-mass vector, then after each iteration the estimator $\hat{\beta}_{1,3}$ is also a $\tilde{c}\mu$-mass vector. By (B.1), each entry of $\hat{\beta}_{1,3}$ can be simplified as,

$$\begin{aligned}
\tilde{\beta}_{1,3,l} = & \frac{w^*_1 \sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\hat{w}_1 \sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \beta^*_{1,3,l} + \frac{w^*_1}{\hat{w}_1} A^{-1}_{ll} B_{ll} \beta^*_{1,3,l} \\
& - \frac{w^*_1}{\hat{w}_1} \frac{\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,1}/n} \langle \beta^*_{1,1}, \hat{\beta}_{1,1}\rangle \langle \beta^*_{1,2}, \hat{\beta}_{1,2}\rangle \beta^*_{1,3,l} \\
& + \frac{1}{\hat{w}_1} A^{-1}_{ll} \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,1} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\mathcal{E}_{iF})_{l_1,l_2,l} \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2},
\end{aligned}$$

and then

$$\begin{aligned}
|\tilde{\beta}_{1,3,l}| &\leq \frac{|w^*_1 B_{ll}|\mu}{|\hat{w}_1 A_{ll}|\sqrt{s}} + \frac{\left| \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,1} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\mathcal{E}_{iF})_{l_1,l_2,l} \hat{\beta}_{1,1,l_1} \hat{\beta}_{1,2,l_2} \right|}{|\hat{w}_1 A_{ll}|} \\
&\leq 2 \frac{\frac{1}{n}\sum_{i=1}^n \hat{\alpha}_{i,1}\alpha^*_{i,1} + \gamma}{\frac{1}{n}\sum_{i=1}^n \hat{\alpha}^2_{i,1} - \gamma} \frac{\mu}{\sqrt{s}} + \left\| \frac{1}{\hat{w}_1} A^{-1} \left\{ \sum_{i=1}^n \frac{\hat{\alpha}_{i,1}}{n} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{1,1} \times_2 \hat{\beta}_{1,2} \right\} \right\|
\end{aligned}$$

Then, together with (A.2) and Proposition A.12, we get

$$|\tilde{\beta}_{1,3,l}| \leq 2 \frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma} \frac{\mu}{\sqrt{s}} + \frac{2\tilde{C}\sigma}{w^*_1 \lambda_{\min}} \sqrt{\frac{s\log(d)}{np}} + 6\gamma'\epsilon,$$

for some large enough $\tilde{C} > 0$. Note that γ' is arbitrary small. Also, by Assumption 2(c), $\epsilon < \min\left\{ \frac{\lambda^3_{\min}}{24\sqrt{10}c_2\lambda^2_{\max}}, \frac{1}{6} \right\}$, ϵ is arbitrary small. Therefore, we can assume that

$$|\tilde{\beta}_{1,3,l}| \leq 2 \frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma} \frac{\mu}{\sqrt{s}} + \frac{2\tilde{C}\sigma}{w^*_1 \lambda_{\min}} \sqrt{\frac{s\log(d)}{np}},$$

for some large enough $\tilde{C} > 0$. By (A.5), $\gamma < \lambda_{\min}/2$, then we have

$$2 \frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma} \frac{\mu}{\sqrt{s}} \leq \frac{6\lambda_{\max}\mu}{\lambda_{\min}\sqrt{s}},$$

and by Assumption 2(d) with $\tilde{C}^2 = \lambda_{\min}^2 c_5$, we have $np \geq \tilde{C}^2\sigma^2 s^2 \log(d)/\{w^{*2}\lambda_{\min}^2\}$, then

$$\frac{2\tilde{C}\sigma}{w_1^*\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}} \leq \frac{\mu}{\sqrt{s}}. \tag{B.9}$$

Therefore, there is some global constant (does not depend on iterations) $\tilde{c} > 0$, such that

$$\max_l \left\{|\hat{\beta}_{k,3,l}|\right\} \leq \tilde{c}\frac{\mu}{\sqrt{s}}.$$

Given that the true parameter $\beta_{k,3}^*$ is a μ-mass vector, the update from each iteration is a $\tilde{c}\mu$-mass vector.

Error bound of the estimator from Step 4

Now we can derive the error bound for the estimator $\hat{\beta}_{1,4}$. That is, we aim to bound $\|\hat{\beta}_{1,4} - \beta_{1,4}^*\|$, given the other estimators $\hat{w}_1, \hat{\beta}_{1,1}, \hat{\beta}_{1,2}, \hat{\beta}_{1,3}$.

Denote $\hat{A}_1 = \hat{w}_1\hat{\beta}_{1,1} \circ \hat{\beta}_{1,2} \circ \hat{\beta}_{1,3}$, and $A_1^* = w_1^*\beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^*$.

For the case of $r = 1$, the true model is $Y_i = \beta_{1,4}^{*T} x_i A_1^* + \mathcal{E}_i$. Then, from (8), the closed form solution of $\hat{\beta}_{1,4}$ becomes

$$\hat{\beta}_{1,4} = \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}(\hat{A}_1)\right\|_F^2 x_i x_i^T\right\}^{-1} \frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(\beta_{1,4}^{*T} x_i A_1^* + \mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i.$$

We can write $\left\|\Pi_{\Omega_i}(\hat{A}_1)\right\|_F^2$ as $\langle \Pi_{\Omega_i}(\hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1)\rangle$. Then, we have

$$\hat{\beta}_{1,4} - \beta_{1,4}^* = \left\| \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}(\hat{A}_1)\right\|_F^2 x_i x_i^T\right\}^{-1} \times \left\{\frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A_1^*), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i x_i^T \beta_{1,4}^*\right.\right.$$

$$\left.\left. + \frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i - \frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(\hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i x_i^T \beta_{1,4}^*\right\}.$$

Therefore,

$$\|\hat{\beta}_{1,4} - \beta_{1,4}^*\| = \left\| \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}(\hat{A}_1)\right\|_F^2 x_i x_i^T\right\}^{-1} \times \right.$$

$$\left. \times \left\{\frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A_1^* - \hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i x_i^T \beta_{1,4}^* + \frac{1}{n}\sum_i \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1)\right\rangle x_i\right\}\right\|,$$

and then

$$\|\hat{\beta}_{1,4} - \beta_{1,4}^*\| \leq \underbrace{\left\|\left\{\frac{1}{np}\sum_{i=1}^n \left\|\Pi_{\Omega_i}(\hat{A}_1)\right\|_F^2 x_i x_i^T\right\}^{-1}\right\|}_{I_2} \times$$

$$\Bigg(\underbrace{\left\| \frac{1}{np} \sum_{i=1}^{n} \left\langle \Pi_{\Omega_i}(A_1^* - \hat{A}_1), \Pi_{\Omega_i}(\hat{A}_1) \right\rangle x_i x_i^T \beta_{1,4}^* \right\|}_{II_2} + \underbrace{\left\| \frac{1}{np} \sum_i \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_1) \right\rangle x_i \right\|}_{III_2} \Bigg).$$

Combining the bounds of I_2, II_2, III_2 from Propositions A.14 - A.16, with probability at least $1 - 1/d^9$, we obtain that,

$$\|\hat{\beta}_{1,4} - \beta_{1,4}^*\| \leq \frac{8}{\lambda_{\min} w_1^{*2}} \left[\{6c_2 + \gamma\} w_1^{*2} \epsilon + \frac{3\tilde{C}_2 \sigma w_1^*}{2} \sqrt{\frac{qs\log(d)}{np}} \right].$$

Note that, from (A.5) $\gamma < c_2$. Then, by letting $k_2 = \frac{56c_2}{\lambda_{\min}}$ and defining $\tilde{C}_2^* = 12\tilde{C}_2$ since $\tilde{C}_2$ is any constant, we conclude that, with probability at least $1 - 1/d^9$

$$\|\hat{\beta}_{1,4} - \beta_{1,4}^*\| \leq k_2 \epsilon + \frac{\tilde{C}_2^* \sigma}{\lambda_{\min} w_1^*} \sqrt{\frac{qs\log(d)}{np}}. \tag{B.10}$$

Next, we combine the steps to prove Theorem 5.1.

We iteratively apply the error bound from each step, and obtain the final error bound in Theorem 5.1.

Given the initial estimators $\hat{\beta}_{1,j}^{(0)}$ and $\hat{w}_1^{(0)}$ with an initialization error ϵ, the error bound in (B.8) implies that, with probability at least $1 - 1/d^9$,

$$\|\hat{\beta}_{1,3}^{(1)} - \beta_{1,3}^*\| \leq k_1 \epsilon + \frac{6\sqrt{10}\lambda_{\max} \tilde{C} \sigma}{w_1^* \lambda_{\min}^2} \sqrt{\frac{s\log(d)}{np}},$$

where

$$k_1 = \frac{6\sqrt{10}\lambda_{\max}}{\lambda_{\min}} \left\{ \frac{\lambda_{\max}}{\lambda_{\min}} + \frac{2\gamma}{\lambda_{\min}} + 3\gamma' \right\} = \frac{6\sqrt{10}\lambda_{\max}^2 \epsilon}{\lambda_{\min}^2} + \frac{12\sqrt{10}\lambda_{\max}\gamma}{\lambda_{\min}^2} + \frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}},$$

defining γ' as

$$\gamma' = \frac{1}{2} \min \left\{ \frac{\lambda_{\min}}{72\sqrt{10}\lambda_{\max}}, \frac{\lambda_{\min}^2}{144\sqrt{10}\lambda_{\max} c_2} \right\}. \tag{B.11}$$

Then, we have that $\frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}} < \frac{1}{4}$. By (A.5) we have $\gamma < \frac{\lambda_{\min}{}^2}{96\sqrt{10}\lambda_{\max}}$, which gives $\frac{12\sqrt{10}\lambda_{\max}\gamma}{\lambda_{\min}^2} < \frac{1}{8} < \frac{1}{4}$. Also, using Assumption 2(c), we have $\frac{6\sqrt{10}\lambda_{\max}^2\epsilon}{\lambda_{\min}^2} < \frac{1}{4}$.
Therefore, $k_1 < 1$. Similarly, the error bound holds for $\|\hat{\beta}_{1,1}^{(1)} - \beta_{1,1}^*\|$, $\|\hat{\beta}_{1,2}^{(1)} - \beta_{1,2}^*\|$ and $|\hat{w}_1^{(1)} - w_1^*|/w_1^*$.

By (B.10), with probability at least $1 - 2/d^9$, we have that

$$\|\hat{\beta}_{1,4} - \beta_{1,4}^*\| \leq k_2 \left(k_1 \epsilon + \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w_1^* \lambda_{\min}^2} \sqrt{\frac{s\log(d)}{np}} \right) + \frac{\tilde{C}_2^* \sigma}{\lambda_{\min} w_1^*} \sqrt{\frac{qs\log(d)}{np}},$$

and then

$$\|\hat{\beta}_{1,4} - \beta^*_{1,4}\| \leq k_2 k_1 \epsilon + k_2 \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w_1^* \lambda_{\min}^2}\sqrt{\frac{s\log(d)}{np}} + \frac{\tilde{C}_2\sigma}{\lambda_{\min} w_1^*}\sqrt{\frac{qs\log(d)}{np}}.$$

The contraction coefficient is

$$k = k_1 k_2 = \left\{ \frac{6\sqrt{10}\lambda_{\max}^2 \epsilon}{\lambda_{\min}^2} + \frac{12\sqrt{10}\lambda_{\max}\gamma}{\lambda_{\min}^2} + \frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}} \right\} \frac{c_2}{\lambda_{\min}}.$$

We have that $\frac{18\sqrt{10}\lambda_{\max}c_2\gamma'}{\lambda_{\min}^2} < \frac{1}{4}$. By (A.5), we have $\frac{12\sqrt{10}\lambda_{\max}c_2\gamma}{\lambda_{\min}^3} < \frac{1}{4}$. Also, using Assumption 2(c), we have $\frac{6\sqrt{10}c_2\lambda_{\max}^2\epsilon}{\lambda_{\min}^3} < \frac{1}{4}$.
Therefore, $k < 1$.

We have now obtained the error bound from the first iteration. After repeatedly plugging the estimation error bound from iteration $(t-1)$ into the error bound from iteration t, with probability at least $1 - (t+1)/d^9$, we have that

$$\max\{|\hat{w}_1^{(t)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(t)} - \beta_{1,j}^*\|_2\}$$

$$\leq k^t\epsilon + \frac{1-k^t}{1-k}\frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w_1^*\lambda_{\min}^2}\sqrt{\frac{s\log(d)}{np}} + \frac{1-k^{t-1}}{1-k}\frac{\tilde{C}_2\sigma}{\lambda_{\min}w_1^*}\sqrt{\frac{qs\log(d)}{np}}.$$

Then,

$$\max\{|\hat{w}_1^{(t)} - w_1^*|/w_1^*, \max_j \|\hat{\beta}_{1,j}^{(t)} - \beta_{1,j}^*\|_2\} \leq k^t\epsilon + \frac{1}{1-k}\frac{C_1\sigma}{w_1^*}\sqrt{\frac{s\log(d)}{np}},$$

where $C_1 = (6\sqrt{10}c_2\lambda_{\max}^2 + \tilde{C}_2\lambda_{\min}\sqrt{q})/\lambda_{\min}^2$.

This completes the proof of Theorem 5.1. □

B.2 Proof of Theorem 5.2

We follow the similar stpes as for the proof of Theorem 5.1. Thus, the first step is to bound the error for the unconstrained estimator from Step 1 of Algorithm 1.

The model for a general rank r is of the form:

$$Y_i = \sum_{k\in[r]} \beta_{k,4}^{*T} x_i w_k^* \beta_{k,1}^* \circ \beta_{k,2}^* \circ \beta_{k,3}^* + \mathcal{E}_i, i = 1, \ldots, n.$$

Error bound of the estimator from Step 1

In (6) we showed that the closed form solution of the optimization problem (4) for $\tilde{\beta}_{1,3}$ is:

$$\tilde{\beta}_{k,3,l} = \frac{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,k}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{R}_{i,1,l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2}}{\sum_{i=1}^{n} \frac{1}{n}(\hat{\alpha}_{i,k})^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{w}_k (\hat{\beta}_{k,1,l_1})^2 (\hat{\beta}_{k,2,l_2})^2},$$

where $\hat{R}_i = \left(Y_i - \sum_{k' \neq k} \hat{w}_{k'} \hat{\alpha}_{i,k} \hat{\beta}_{k',1} \circ \hat{\beta}_{k',2} \circ \hat{\beta}_{k',3} \right) / \hat{\alpha}_{i,k}$ and $\hat{\alpha}_{i,k} = \hat{\beta}_{k,4}^T x_i$.

Denote $F_1 = \text{supp}(\beta_{k,1}^*) \cup \text{supp}(\hat{\beta}_{k,1}^*)$, $F_2 = \text{supp}(\beta_{k,2}^*) \cup \text{supp}(\hat{\beta}_{k,2}^*)$, $F_3 = \text{supp}(\beta_{k,3}^*) \cup \text{supp}(\hat{\beta}_{k,3}^*)$, where $\text{supp}(v)$ refers to the set of indices in v that are nonzero. Let $F = F_1 \cup F_2 \cup F_3$. Then, we consider the following estimator, that is equivalent to the estimator above.

$$\tilde{\beta}_{k,3,l} = \frac{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,k}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\hat{R}_{iF})_{l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2}}{\sum_{i=1}^{n} \frac{1}{n}\hat{\alpha}_{i,k}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} \hat{w}_k (\hat{\beta}_{k,1,l_1})^2 (\hat{\beta}_{k,2,l_2})^2}.$$

For the vector $\beta_{k,j}$, denote $\bar{\beta}_{k,j} = \textit{Truncate}(\beta_{k,j}^*, F_j)$, for $k = 1, \ldots, r$ and $j = 1, 2, 3$. By definition of F_j, we have $\bar{\beta}_{k,j}^* = \beta_{k,j}^*$ and $\bar{\hat{\beta}}_{k,j} = \hat{\beta}_{k,j}$, for $j = 1, 2, 3$. By Lemma A.7 and substituting the expression of $\hat{R}_i$ into $\tilde{\beta}_{k,3,l}$, the vector $\tilde{\beta}_{k,3}$ can be expanded as

$$\begin{aligned}
\tilde{\beta}_{k,3} = & \frac{w_k^* \sum_{i=1}^{n} \hat{\alpha}_{i,k} \alpha_{i,k}^* / n}{\hat{w}_k \sum_{i=1}^{n} \hat{\alpha}_{i,k}^2 / n} \langle \beta_{k,1}^*, \hat{\beta}_{k,1} \rangle \langle \beta_{k,2}^*, \hat{\beta}_{k,2} \rangle \beta_{k,3}^* \\
& + \frac{\sum_{k' \neq k} \left\{ \sum_{i=1}^{n} \frac{1}{n} \hat{\alpha}_{i,k} \alpha_{i,k'}^* \right\} w_{k'}^* \langle \bar{\beta}_{k',1}^*, \hat{\beta}_{k,1} \rangle \langle \bar{\beta}_{k',2}^*, \hat{\beta}_{k,2} \rangle \bar{\beta}_{k',3}^*}{\hat{w}_k \sum_{i=1}^{n} \hat{\alpha}_{i,k}^2 / n} \\
& - \frac{\sum_{k' \neq k} \left\{ \sum_{i=1}^{n} \frac{1}{n} \hat{\alpha}_{i,k} \hat{\alpha}_{i,k'} \right\} \hat{w}_{k'}^* \langle \bar{\hat{\beta}}_{k',1}^*, \hat{\beta}_{k,1} \rangle \langle \bar{\hat{\beta}}_{k',2}^*, \hat{\beta}_{k,2} \rangle \bar{\hat{\beta}}_{k',3}^*}{\hat{w}_k \sum_{i=1}^{n} \hat{\alpha}_{i,k}^2 / n} \\
& + \frac{w_k^*}{\hat{w}_k} A^{-1} \left\{ B - A \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,k} \alpha_{i,k}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,k}^2 / n} \langle \beta_{k,1}^*, \hat{\beta}_{k,1} \rangle \langle \beta_{k,2}^*, \hat{\beta}_{k,2} \rangle \right\} \beta_{k,3}^* \\
& + \sum_{k' \neq k} \frac{w_k^*}{\hat{w}_k} A^{-1} \left\{ F_{k'} - A \frac{\sum_{i=1}^{n} \hat{\alpha}_{i,k} \alpha_{i,k'}^* / n}{\sum_{i=1}^{n} \hat{\alpha}_{i,k}^2 / n} \langle \bar{\beta}_{k',1}^*, \hat{\beta}_{k,1} \rangle \langle \bar{\beta}_{k',2}^*, \hat{\beta}_{k,2} \rangle \right\} \bar{\beta}_{k',3}^*
\end{aligned}$$

$$-\sum_{k'\neq k}\frac{\hat{w}_{k'}}{\hat{w}_k}A^{-1}\left\{G_{k'}-A\frac{\sum_{i=1}^{n}\hat{\alpha}_{i,k}\hat{\alpha}'_{i,k}/n}{\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\right\}\bar{\hat{\beta}}^*_{k',3}$$

$$+\frac{1}{\hat{w}_k}A^{-1}\left\{\sum_{i=1}^{n}\frac{\hat{\alpha}_{i,k}}{n}\Pi_{\Omega_i}(\mathcal{E}_{iF})\times_1\hat{\beta}_{k,1}\times_2\hat{\beta}_{k,2}\right\}, \tag{B.12}$$

where $A, B, F_{k'}$ and $G_{k'}$ are diagonal matrices with diagonal entry,

$$A_{ll}=\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}^2_{k,1,l_1}\hat{\beta}^2_{k,2,l_2}$$

$$B_{ll}=\sum_{i=1}^{n}\hat{\alpha}_{i,k}\alpha^*_{i,k}/n\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\beta^*_{k,1,l_1}\beta^*_{k,2,l_2}$$

$$F_{k'll}=\sum_{i=1}^{n}\frac{1}{n}\hat{\alpha}_{i,k}\alpha^*_{i,k'}\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\bar{\beta}^*_{k',1,l_1}\bar{\beta}^*_{k',2,l_2}$$

$$G_{k'll}=\sum_{i=1}^{n}\frac{1}{n}\hat{\alpha}_{i,k}\hat{\alpha}_{i,k'}\sum_{l_1,l_2}\delta_{i,l_1,l_2,l}\hat{\beta}_{k,1,l_1}\hat{\beta}_{k,2,l_2}\bar{\hat{\beta}}^*_{k',1,l_1}\bar{\hat{\beta}}^*_{k',2,l_2}.$$

Then, the difference between $\tilde{\beta}_{k,3}$ and $\frac{w_k^*\sum_{i=1}^{n}\hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\beta^*_{k,3}$ is

$$\tilde{\beta}_{k,3}-\frac{w_k^*\sum_{i=1}^{n}\hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\beta^*_{k,3}=I_1+I_2+I_3+I_4+I_5, \tag{B.13}$$

where

$$I_1=\frac{w_k^*\sum_{i=1}^{n}\hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\left\{\langle\beta^*_{k,1},\hat{\beta}_{k,1}\rangle\langle\beta^*_{k,2},\hat{\beta}_{k,2}\rangle-1\right\}\beta^*_{k,3}$$

$$II_1=\frac{1}{\hat{w}_k\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\sum_{i=1}^{n}\frac{1}{n}\hat{\alpha}_{i,k}\sum_{k'\neq k}\left\{\alpha^*_{i,k'}w^*_{k',1}\langle\bar{\beta}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\beta}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\beta}^*_{k',3}\right.$$

$$\left.-\hat{\alpha}_{i,k'}\hat{w}^*_{k',1}\langle\bar{\hat{\beta}}^*_{k',1},\hat{\beta}_{k,1}\rangle\langle\bar{\hat{\beta}}^*_{k',2},\hat{\beta}_{k,2}\rangle\bar{\hat{\beta}}^*_{k',3}\right\}$$

$$III_1=\frac{w_k^*}{\hat{w}_k}A^{-1}\left\{B-A\frac{\sum_{i=1}^{n}\hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\sum_{i=1}^{n}\hat{\alpha}_{i,k}^2/n}\langle\beta^*_{k,1},\hat{\beta}_{k,1}\rangle\langle\beta^*_{k,2},\hat{\beta}_{k,2}\rangle\right\}\beta^*_{k,3}$$

$$IV_1 = \sum_{k' \neq k} \frac{w^*_{k'}}{\hat{w}_k} A^{-1} \left\{ F_{k'} - A \frac{\sum_{i=1}^n \hat{\alpha}_{i,k} \alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1} \rangle \langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2} \rangle \right\} \bar{\beta}^*_{k',3}$$

$$- \sum_{k' \neq k} \frac{\hat{w}^*_{k'}}{\hat{w}_k} A^{-1} \left\{ G_{k'} - A \frac{\sum_{i=1}^n \hat{\alpha}_{i,k} \hat{\alpha}_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \langle \bar{\hat{\beta}}^*_{k',1}, \hat{\beta}_{k,1} \rangle \langle \bar{\hat{\beta}}^*_{k',2}, \hat{\beta}_{k,2} \rangle \right\} \bar{\hat{\beta}}^*_{k',3}$$

$$V_1 = \frac{1}{\hat{w}_k} A^{-1} \left\{ \sum_{i=1}^n \frac{\hat{\alpha}_{i,k}}{n} \Pi_{\Omega_i}(\mathcal{E}_{iF}) \times_1 \hat{\beta}_{k,1} \times_2 \hat{\beta}_{k,2} \right\}.$$

Comparing (B.13) with (B.2) in the rank-1 case, we see that, when $r = 1$, the sum includes the terms I_1, III_1 and V_1. When the rank $r > 1$, the sum includes two additional terms II_1 and IV_1, which appear due to the interplay among different ranks. By Proposition A.5, we have

$$\|I_1\| \leq \frac{2\lambda_{\max}}{\lambda_{\min}} \epsilon^2.$$

By Proposition A.8, with probability at least $1 - 4/d^{10}$, we have

$$\|III_1\| \leq \frac{4\gamma}{\lambda_{\min}} \epsilon,$$

where γ is a positive constant equal to

$$\gamma = \frac{1}{2} \min \left\{ \frac{\lambda_{\min}}{2}, c_2, \frac{\lambda^2_{\min}}{192\sqrt{10}\lambda_{\max}}, \frac{\lambda^3_{\min} w^{*2}_{\min}}{96\sqrt{10} c_2 \lambda_{\max} w^{*2}_{\max} r} \right\}. \tag{B.14}$$

Furthermore, by Proposition A.12, with probability at least $1 - 10/d^{10}$, we have,

$$\|V_1\| \leq \frac{2\tilde{C}\sigma}{\lambda_{\min} w^*_{\min}} \sqrt{\frac{s\log(d)}{np}} + 6\gamma'\epsilon.$$

The value of γ' will be determined later. Next, by Proposition A.17, we have

$$\|II_1\| \leq \frac{c_1^2 r w^*_{\max} \{8\epsilon^2 + 8\xi\epsilon + 5\xi^2\epsilon\}}{\lambda_{\min} w^*_{\min}}.$$

Proposition A.18 gives that with probability at least $1 - 2/d^{10}$,

$$\|IV_1\| \leq \frac{4\gamma\epsilon}{\lambda_{\min}}.$$

Hence, we can bound the term in (B.13). We have

$$\left\| \tilde{\beta}_{k,3} - \frac{w^*_k \sum_{i=1}^n \hat{\alpha}_{i,k} \alpha^*_{i,k}/n}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n} \beta^*_{k,3} \right\| \leq \|I_1\| + \|II_1\| + \|III_1\| + \|IV_1\| + \|V_1\|$$

$$\leq \frac{2\lambda_{\max}}{\lambda_{\min}}\epsilon^2 + \frac{c_1^2 r w^*_{\max}\{8\epsilon^2 + 8\xi\epsilon + 5\xi^2\epsilon\}}{\lambda_{\min} w^*_{\min}} + \frac{8\gamma}{\lambda_{\min}}\epsilon + 6\gamma'\epsilon + \frac{2\tilde{C}\sigma}{w^*_{\min}\lambda_{\min}}\sqrt{\frac{s\log(d)}{np}}.$$

As shown in the rank-1 case in (B.5), the error of normalized $\tilde{\beta}_{k,3}$ can be bounded as

$$\left|\frac{w_k^* \sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\right| \left\|\frac{\tilde{\beta}_{k,3}}{\|\tilde{\beta}_{k,3}\|} - \beta^*_{k,3}\right\| \leq 2\left\|\tilde{\beta}_{k,3} - \frac{w_k^* \sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\beta^*_{k,3}\right\|.$$

Recall from (A.1) that $|\hat{w}_k - w_k^*| \leq \frac{1}{2}w_k^*$. Then, $\frac{\hat{w}_k}{w_k^*} \leq \frac{3}{2}$. Using (A.2), we have

$$\left|\frac{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}{w_k^* \sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k}/n}\right| \leq \frac{3\lambda_{\max}}{2\lambda_{\min}}.$$

Therefore, we have

$$\left\|\frac{\tilde{\beta}_{k,3}}{\|\tilde{\beta}_{k,3}\|} - \beta^*_{k,3}\right\| \leq \frac{6\lambda^2_{\max}}{\lambda^2_{\min}}\epsilon^2 + \frac{3c_1^2\lambda_{\max} r w^*_{\max}\{8\epsilon^2 + 8\xi\epsilon + 5\xi^2\epsilon\}}{\lambda^2_{\min} w^*_{\min}} + \frac{24\lambda_{\max}\gamma}{\lambda^2_{\min}}\epsilon$$
$$+\frac{18\gamma'\lambda_{\max}\epsilon}{\lambda_{\min}} + \frac{6\tilde{C}\lambda_{\max}\sigma}{w^*_{\min}\lambda^2_{\min}}\sqrt{\frac{s\log(d)}{np}}.$$

Error bound of the estimator from Step 2

In this step we derive the error bound for the unconstrained estimator from Step 2 for the general case. Similarly, to the case for $r = 1$, from (B.7) we have

$$\|\bar{\beta}_{k,3} - \beta^*_{k,3}\| \leq \sqrt{10}\left\|\frac{\tilde{\beta}_{k,3}}{\|\tilde{\beta}_{k,3}\|} - \beta^*_{k,3}\right\| \leq k_1\epsilon + \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w^*_{\min}\lambda^2_{\min}}\sqrt{\frac{s\log(d)}{np}},$$

where

$$k_1 = \frac{6\sqrt{10}\lambda^2_{\max}\epsilon}{\lambda^2_{\min}} + \frac{24\sqrt{10}\lambda_{\max}\gamma}{\lambda^2_{\min}} + \frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}} + \frac{c_1^2\lambda_{\max} r w^*_{\max}}{\lambda^2_{\min} w^*_{\min}}\epsilon$$
$$+\frac{c_1^2\lambda_{\max} r w^*_{\max}}{\lambda^2_{\min} w^*_{\min}}\xi.$$

By (B.14) $\gamma < \frac{\lambda^2_{\min}}{192\sqrt{10}\lambda_{\max}}$, then $\frac{24\sqrt{10}\lambda_{\max}\gamma}{\lambda^2_{\min}} < \frac{1}{8}$. Setting constant $\gamma' < \frac{\lambda_{\min}}{144\sqrt{10}\lambda_{\max}}$, we have $\frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}} < \frac{1}{8}$. By Assumption 3(c), we have that $\epsilon < \frac{\lambda^2_{\min}}{\lambda^2_{\max}}$, then $\frac{6\sqrt{10}\lambda^2_{\max}\epsilon}{\lambda^2_{\min}} < \frac{1}{4}$. Also, by Assumption 3(c), $\frac{c_1^2\lambda_{\max} r w^*_{\max}}{\lambda^2_{\min} w^*_{\min}}\epsilon < \frac{1}{4}$. By Assumption 3(d), we have that $\frac{c_1^2\lambda_{\max} r w^*_{\max}}{\lambda^2_{\min} w^*_{\min}}\xi < \frac{1}{4}$.

Therefore, $k_1 < 1$. Finally, we prove that the estimator $\hat{\beta}_{k,3}$ has the μ-mass property under the assumption that the true parameter $\beta^*_{k,3}$ is a μ-mass vector.

Using (B.12), each entry of vector $\tilde{\beta}_{k,3}$ can be simplified as

$$\tilde{\beta}_{k,3,l} = \underbrace{\frac{w^*_k \sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\langle \beta^*_{k,1}, \hat{\beta}_{k,1}\rangle\langle \beta^*_{k,2}, \hat{\beta}_{k,2}\rangle \beta^*_{k,3,l}}_{(*)}$$

$$+ \underbrace{\frac{\sum_{k'\neq k}\left\{\sum_{i=1}^n \frac{1}{n}\hat{\alpha}_{i,k}\alpha^*_{i,k'}\right\} w^*_{k'}\langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1}\rangle\langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2}\rangle \bar{\beta}^*_{k',3,l}}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}}_{(**)}$$

$$\underbrace{- \frac{\sum_{k'\neq k}\left\{\sum_{i=1}^n \frac{1}{n}\hat{\alpha}_{i,k}\hat{\alpha}_{i,k'}\right\} \hat{w}^*_{k'}\langle \bar{\hat{\beta}}^*_{k',1}, \hat{\beta}_{k,1}\rangle\langle \bar{\hat{\beta}}^*_{k',2}, \hat{\beta}_{k,2}\rangle \bar{\hat{\beta}}^*_{k',3,l}}{\hat{w}_k \sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}}_{(***)} + \frac{w^*_k}{\hat{w}_k} A^{-1}_{ll} B_{ll} \beta^*_{k,3,l}$$

$$\underbrace{- \frac{w^*_k}{\hat{w}_k} \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\langle \beta^*_{k,1}, \hat{\beta}_{k,1}\rangle\langle \beta^*_{k,2}, \hat{\beta}_{k,2}\rangle \beta^*_{k,3,l}}_{(*)} + \sum_{k'\neq k} \frac{w^*_k}{\hat{w}_k} A^{-1}_{ll} F_{k'll} \bar{\beta}^*_{k',3,l}$$

$$\underbrace{- \sum_{k'\neq k} \frac{w^*_k}{\hat{w}_k} \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\alpha^*_{i,k'}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\langle \bar{\beta}^*_{k',1}, \hat{\beta}_{k,1}\rangle\langle \bar{\beta}^*_{k',2}, \hat{\beta}_{k,2}\rangle \bar{\beta}^*_{k',3,l}}_{(**)} - \sum_{k'\neq k} \frac{\hat{w}_{k'}}{\hat{w}_k} A^{-1}_{ll} G_{k'll} \bar{\hat{\beta}}^*_{k',3,l}$$

$$\underbrace{+ \sum_{k'\neq k} \frac{\hat{w}_{k'}}{\hat{w}_k} \frac{\sum_{i=1}^n \hat{\alpha}_{i,k}\hat{\alpha}'_{i,k}/n}{\sum_{i=1}^n \hat{\alpha}^2_{i,k}/n}\langle \bar{\hat{\beta}}^*_{k',1}, \hat{\beta}_{k,1}\rangle\langle \bar{\hat{\beta}}^*_{k',2}, \hat{\beta}_{k,2}\rangle \bar{\hat{\beta}}^*_{k',3,l}}_{(***)}$$

$$+ \frac{1}{\hat{w}_k} A^{-1}_{ll} \left\{ \sum_{i=1}^n \frac{1}{n}\hat{\alpha}^2_{i,k} \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\mathcal{E}_{iF})_{l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \right\},$$

and then,

$$\tilde{\beta}_{k,3,l} = \frac{w_k^*}{\hat{w}_k} A_{ll}^{-1} B_{ll} \beta_{k,3,l}^* + \sum_{k' \neq k} \frac{w_k^*}{\hat{w}_k} A_{ll}^{-1} F_{k'll} \bar{\beta}_{k',3,l}^* - \sum_{k' \neq k} \frac{\hat{w}_{k'}}{\hat{w}_k} A_{ll}^{-1} G_{k'll} \bar{\hat{\beta}}_{k',3,l}^*$$

$$+ \frac{1}{\hat{w}_k} A_{ll}^{-1} \left\{ \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,k}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\mathcal{E}_{iF})_{l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \right\}.$$

Therefore,

$$|\tilde{\beta}_{k,3,l}| \leq \frac{|w_k^* B_{ll}| \mu}{|\hat{w}_k A_{ll}| \sqrt{s}} + \frac{\left| \sum_{i=1}^n \frac{1}{n} \hat{\alpha}_{i,k}^2 \sum_{l_1,l_2} \delta_{i,l_1,l_2,l} (\mathcal{E}_{iF})_{l_1,l_2,l} \hat{\beta}_{k,1,l_1} \hat{\beta}_{k,2,l_2} \right|}{|\hat{w}_k A_{ll}|}$$

$$+ \sum_{k' \neq k} \left(\frac{|w_{k'}^* F_{k'll}| \mu}{|\hat{w}_k A_{ll}| \sqrt{s}} + \frac{|\hat{w}_{k'}^* G_{k'll}| \mu}{|\hat{w}_k A_{ll}| \sqrt{s}} \right).$$

Then

$$|\tilde{\beta}_{k,3,l}| \leq 2 \frac{\frac{1}{n} \sum_{i=1}^n \hat{\alpha}_{i,k} \alpha_{i,k}^* + \gamma}{\frac{1}{n} \sum_{i=1}^n \hat{\alpha}_{i,k}^2 - \gamma} \frac{\mu}{\sqrt{s}} + \sum_{k'=1, k' \neq k}^r \left\{ \frac{2 w_{\max}^*}{w_{\min}^*} \frac{\frac{1}{n} \sum_{i=1}^n \hat{\alpha}_{i,k} \alpha_{i,k'}^* + \gamma}{\frac{1}{n\epsilon} \sum_{i=1}^n \hat{\alpha}_{i,k}^2 - \gamma} \frac{\mu}{\sqrt{s}} \right.$$

$$\left. + \frac{3 w_{\max}^*}{w_{\min}^*} \frac{\frac{1}{n} \sum_{i=1}^n \hat{\alpha}_{i,k} \hat{\alpha}_{i,k'} + \gamma}{\frac{1}{n\epsilon} \sum_{i=1}^n \hat{\alpha}_{i,k}^2 - \gamma} \frac{\mu}{\sqrt{s}} \right\} + \left\| \frac{1}{\hat{w}_k} A^{-1} \sum_{i=1}^n \frac{\hat{\alpha}_{i,k}}{n} \Pi_{\Omega_i} (\mathcal{E}_{iF}) \times_1 \hat{\beta}_{k,1} \times_2 \hat{\beta}_{k,2} \right\|.$$

Therefore, by using Proposition A.12 and (A.2), we have

$$|\tilde{\beta}_{k,3,l}| \leq 2 \frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma} \frac{\mu}{\sqrt{s}} + \frac{5(r-1) w_{\max}^*}{w_{\min}^*} \frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma} \frac{\mu}{\sqrt{s}} + \frac{2 \tilde{C} \sigma}{w_{\min}^* \lambda_{\min}} \sqrt{\frac{s \log(d)}{np}} + 6 \gamma' \epsilon,$$

for some large enough $\tilde{C} > 0$. Then, by (B.14) $\gamma < \lambda_{\min}/2$, we get

$$|\tilde{\beta}_{k,3,l}| \leq 4 \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu}{\sqrt{s}} + \frac{10(r-1) w_{\max}^*}{w_{\min}^*} \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu}{\sqrt{s}} + \frac{2 \tilde{C} \sigma}{w_{\min}^* \lambda_{\min}} \sqrt{\frac{s \log(d)}{np}} + 6 \gamma' \epsilon,$$

for some large enough $\tilde{C} > 0$. Note that γ' is arbitrary small. Also, by Assumption 3(c), $\epsilon < \min \left\{ \frac{\lambda_{\min}^3 w_{\min}^{*2}}{24\sqrt{10} c_2 \lambda_{\max}^2 w_{\max}^{*2} r}, \frac{\lambda_{\min}^3 w_{\min}^{*3}}{4 c_2 \lambda_{\max} w_{\max}^{*3} r^2}, \frac{1}{6} \right\}$, ϵ is arbitrary small. Therefore, we can assume that

$$|\tilde{\beta}_{1,3,l}| \leq 4 \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu}{\sqrt{s}} + \frac{10(r-1) w_{\max}^*}{w_{\min}^*} \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mu}{\sqrt{s}} + \frac{2 \tilde{C} \sigma}{w_{\min}^* \lambda_{\min}} \sqrt{\frac{s \log(d)}{np}}.$$

for some large enough $\tilde{C} > 0$. Similarly to the proof for $r = 1$ in (B.9), under Assumption 3(e), we have

$$\frac{\tilde{C}\sigma}{w^*_{\min}\{\lambda_{\min} - \gamma\}}\sqrt{\frac{s\log(d)}{np}} \leq \frac{\mu}{\sqrt{s}},$$

and by (B.14) $\gamma < \lambda_{\min}/2$, we have

$$2\frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma}\frac{\mu}{\sqrt{s}} \leq 4\frac{\lambda_{\max}}{\lambda_{\min}}\frac{\mu}{\sqrt{s}} \text{ and}$$

$$\frac{5(r-1)w^*_{\max}}{w^*_{\min}}\frac{\lambda_{\max} + \gamma}{\lambda_{\min} - \gamma}\frac{\mu}{\sqrt{s}} \leq \frac{10(r-1)w^*_{\max}\lambda_{\max}}{w^*_{\min}\lambda_{\min}}\frac{\mu}{\sqrt{s}}.$$

Therefore, for some constant $\tilde{c}$, we have that

$$\max_l \left\{|\hat{\beta}_{k,3,l}|\right\} \leq \tilde{c}\frac{\mu}{\sqrt{s}}.$$

Error bound of the estimator from Step 4

In the third step, we bound $\|\hat{\beta}_{k,4} - \beta^*_{k,4}\|$ for each k given all other parameters $\hat{w}_k, \hat{\beta}_{k,1}, \hat{\beta}_{k,2}, \hat{\beta}_{k,3}$ and $\hat{w}_{k'}, \hat{\beta}_{k',1}, \hat{\beta}_{k',2}, \hat{\beta}_{k',3}$ for $k' \neq k$. The closed form solution of the estimator will be

$$\hat{\beta}_{k,4} = \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}\left(\hat{A}_k\right)\right\|_F^2 x_i x_i^T\right\}^{-1} \frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(\hat{S}_{ik}), \hat{A}_k\right\rangle x_i, \tag{B.15}$$

where $\hat{S}_{i,k} = Y_i - \sum_{k' \neq k} \hat{w}_{k'}\hat{\beta}^T_{k',4}x_i\hat{\beta}_{k',1} \circ \hat{\beta}_{k',2} \circ \hat{\beta}_{k',3}$ and $\hat{A}_k = \hat{w}_k\hat{\beta}_{k,1} \circ \hat{\beta}_{k,2} \circ \hat{\beta}_{k,3}$.

Plugging Y_i into (B.15), $\hat{\beta}_{k,4}$ can be written as

$$\hat{\beta}_{k,4} = \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}\left(\hat{A}_k\right)\right\|_F^2 x_i x_i^T\right\}^{-1}$$

$$\times\frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}\left(\sum_{k=1}^r A^*_k x_i^T\beta^*_{k,4} + \mathcal{E}_i - \sum_{k' \neq k}\hat{A}_{k'}x_i^T\hat{\beta}_{k',4}\right), \hat{A}_k\right\rangle x_i$$

$$= \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}\left(\hat{A}_k\right)\right\|_F^2 x_i x_i^T\right\}^{-1} \left(\frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A^*_k), \hat{A}_k\right\rangle x_i x_i^T\beta^*_{k,4}\right.$$

$$\left. + \frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_k\right\rangle x_i + \frac{1}{n}\sum_i\sum_{k' \neq k}\left\langle \Pi_{\Omega_i}(A^*_{k'} - \hat{A}_{k'}), \hat{A}_k\right\rangle x_i x_i^T\beta^*_{k,4}\right).$$

Hence, we get

$$\hat{\beta}_{k,4} - \beta^*_{k,4} = \left\{\frac{1}{n}\sum_{i=1}^n \left\|\Pi_{\Omega_i}\left(\hat{A}_k\right)\right\|_F^2 x_i x_i^T\right\}^{-1} \left(\frac{1}{n}\sum_{i=1}^n \left\langle \Pi_{\Omega_i}(A^*_k), \hat{A}_k\right\rangle x_i x_i^T\beta^*_{k,4}\right.$$

$$-\frac{1}{n}\sum_{i=1}^{n}\left\langle \Pi_{\Omega_i}(\hat{A}_k), \hat{A}_k \right\rangle x_i x_i^T \beta^*_{k,4} + \frac{1}{n}\sum_{i=1}^{n}\left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_k \right\rangle x_i$$
$$+\frac{1}{n}\sum_{i=1}^{n}\sum_{k'=1,k'\neq k}^{r}\left\langle \Pi_{\Omega_i}(A^*_{k'} - \hat{A}_{k'}), \hat{A}_k \right\rangle x_i x_i^T \beta^*_{k,4}\Bigg),$$

and then

$$\hat{\beta}_{k,4} - \beta^*_{k,4} = \left\{\frac{1}{n}\sum_{i=1}^{n}\left\|\Pi_{\Omega_i}\left(\hat{A}_k\right)\right\|_F^2 x_i x_i^T\right\}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left\langle \Pi_{\Omega_i}(A^*_k - \hat{A}_k)), \hat{A}_k \right\rangle x_i x_i^T \beta^*_{k,4}\right.$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n}\left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_k \right\rangle x_i + \frac{1}{n}\sum_{i=1}^{n}\sum_{k'=1,k'\neq k}^{r}\left\langle \Pi_{\Omega_i}(A^*_{k'} - \hat{A}_{k'}), \hat{A}_k \right\rangle x_i x_i^T \beta^*_{k,4}\right).$$

Therefore, using triangle inequality, we have

$$\|\hat{\beta}_{k,4} - \beta^*_{k,4}\| \leq \left\|\left\{\frac{1}{n}\sum_{i=1}^{n}\|\Pi_{\Omega_i}(A_k)\|_F^2 x_i x_i^T\right\}^{-1}\right\|\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_k \right\rangle x_i\right\|\right.$$
$$\left.+\left\|\frac{1}{n}\sum_{i}\sum_{k'=1}^{r}\left\langle \Pi_{\Omega_i}(A^*_{k'} - \hat{A}_{k'}), \hat{A}_k \right\rangle x_i x_i^T \beta^*_{k,4}\right\|\right).$$

Note that since Π_{Ω_i} is an indicator tensor, $\left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \hat{A}_k \right\rangle = \left\langle \Pi_{\Omega_i}(\mathcal{E}_i), \Pi_{\Omega_i}(\hat{A}_k) \right\rangle$. Then, by applying Propositions A.14 - A.16, we obtain that

$$\|\hat{\beta}_{k,4} - \beta^*_{k,4}\| \leq \frac{8}{\lambda_{\min} w^{*2}_{\min}}\left[r\{6c_2 + \gamma\}w^{*2}_{\max}\epsilon + \frac{3\tilde{C}_2\sigma w_{\max}}{2}\sqrt{\frac{qs\log(d)}{np}}\right].$$

Then, using the fact that $\gamma < c_2$ by (B.14),

$$\|\hat{\beta}_{k,4} - \beta^*_{k,4}\| \leq k_2\epsilon + \frac{12\tilde{C}_2\sigma w^*_{\max}}{\lambda_{\min} w^{*2}_{\min}}\sqrt{\frac{qs\log(d)}{np}}, \tag{B.16}$$

where $k_2 = 56c_2 w^{*2}_{\max} r/\lambda_{\min} w^{*2}_{\min}$.

Next, we combine the steps to prove Theorem 5.2.

Finally, we provide the error rate after t iterations, by iteratively applying the error bound from each step. We have shown that, with probability at least $1 - 1/d^9$,

$$\|\bar{\beta}^{(1)}_{k,3} - \beta^*_{k,3}\| \leq k_1\epsilon + \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w^*_{\min}\lambda^2_{\min}}\sqrt{\frac{s\log(d)}{np}}.$$

Combining with the error bound (B.16), we get that with probability at least $1 - 2/d^9$

$$\|\hat{\beta}^{(1)}_{k,4} - \beta^*_{k,4}\| \leq k_2\left(k_1\epsilon + \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w^*_{\min}\lambda^2_{\min}}\sqrt{\frac{s\log(d)}{np}}\right) + \frac{12\tilde{C}_2\sigma w^*_{\max}}{\lambda_{\min} w^{*2}_{\min}}\sqrt{\frac{qs\log(d)}{np}}.$$

This gives

$$\|\hat{\beta}_{k,4}^{(1)} - \beta_{k,4}^*\| \leq k_2 k_1 \epsilon + k_2 \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w_{\min}^* \lambda_{\min}^2} \sqrt{\frac{s\log(d)}{np}} + \frac{12\tilde{C}_2 \sigma w_{\max}^*}{\lambda_{\min} w_{\min}^{*2}} \sqrt{\frac{qs\log(d)}{np}}.$$

The contraction coefficient is

$$k = k_1 k_2 = \left[\frac{6\sqrt{10}\lambda_{\max}^2 \epsilon}{\lambda_{\min}^2} + \frac{24\sqrt{10}\lambda_{\max}\gamma}{\lambda_{\min}^2} + \frac{18\sqrt{10}\lambda_{\max}\gamma'}{\lambda_{\min}} + \frac{c_1 \lambda_{\max} r w_{\max}^*}{\lambda_{\min}^2 w_{\min}^*} \epsilon \right.$$

$$\left. + \frac{c_1 \lambda_{\max} r w_{\max}^*}{\lambda_{\min}^2 w_{\min}^*} \xi \right] \frac{c_2 w_{\max}^{*2} r}{\lambda_{\min} w_{\min}^{*2}}.$$

By Assumptions 3(c) and 3(d), we have

$$\frac{6\sqrt{10}\lambda_{\max}^2 \epsilon}{\lambda_{\min}^2} \frac{c_2 w_{\max}^{*2} r}{\lambda_{\min} w_{\min}^{*2}} \leq \frac{1}{4},$$

$$\frac{c_1 \lambda_{\max} r w_{\max}^*}{\lambda_{\min}^2 w_{\min}^*} \epsilon \frac{c_2 w_{\max}^{*2} r}{\lambda_{\min} w_{\min}^{*2}} \leq \frac{1}{4},$$

$$\frac{c_1 \lambda_{\max} r w_{\max}^*}{\lambda_{\min}^2 w_{\min}^*} \xi \frac{c_2 w_{\max}^{*2} r}{\lambda_{\min} w_{\min}^{*2}} \leq \frac{1}{4}.$$

Also, by (B.14), we have $\gamma < \lambda_{\min}^3 w_{\min}^{*2} / \{96\sqrt{10}\lambda_{\max} c_2 w_{\max}^{*2} r\}$ and

$$\frac{24\sqrt{10}\lambda_{\max}\gamma}{\lambda_{\min}^2} \frac{c_2 w_{\max}^{*2} r}{\lambda_{\min} w_{\min}^{*2}} \leq \frac{1}{4}.$$

Therefore, the contraction coefficient $k < 1$.

The error bound after t iterations is, with probability at least $1 - (t+1)/d^9$

$$\max\{\max_k |\hat{w}_k^{(t)} - w_k^*| / w_k^*, \max_{k,j} \|\hat{\beta}_{k,j}^{(t)} - \beta_{k,j}^*\|_2\}$$

$$\leq k^t \epsilon + \frac{1-k^t}{1-k} \frac{6\sqrt{10}\lambda_{\max}\tilde{C}\sigma}{w_{\min}^* \lambda_{\min}^2} \sqrt{\frac{s\log(d)}{np}} + \frac{1-k^{t-1}}{1-k} \frac{12\tilde{C}_2 \sigma w_{\max}^*}{\lambda_{\min} w_{\min}^{*2}} \sqrt{\frac{qs\log(d)}{np}}$$

$$\leq k^t \epsilon + \frac{1}{1-k} \max\left\{ C_2' \frac{\sigma}{w_{\min}^*} \sqrt{\frac{s\log(d)}{np}}, \frac{C_2'' \sigma w_{\max}^*}{w_{\min}^{*2}} \sqrt{\frac{s\log(d)}{np}} \right\}.$$

Therefore,

$$\max\{\max_k |\hat{w}_k^{(t)} - w_k^*| / w_k^*, \max_{k,j} \|\hat{\beta}_{k,j}^{(t)} - \beta_{k,j}^*\|_2 \leq k^t \epsilon + \frac{C_2}{1-k} \frac{\sigma w_{\max}^*}{w_{\min}^{*2}} \sqrt{\frac{s\log(d)}{np}},$$

where $C_2' = 6\sqrt{10}\lambda_{\max}/\lambda_{\min}^2, C_2'' = 12\tilde{C}_2\sqrt{q}/\lambda_{\min}$ and $C_2 = C_2' + C_2''$.

This completes the proof of Theorem 5.2. □

B.3 Proof of Theorem 5.3

The proof is mostly based on results and steps from Cai et al, 2021 [18]. We divide the proof into three steps.

Step 1:

In step 1, we show that there exists at least one trial $1 \leq r \leq L$ such that $\beta_{1,1}^*$ is the top singular vector of the population version of $\mathcal{T} \times_3 \tilde{g_1^\tau}$. Let $\mathcal{T}^* = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^*$ and $\gamma^{*\tau} = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \langle \beta_{1,3}^*, \tilde{g}_1^\tau \rangle$.

Then

$$\mathcal{T}^* \times_3 \tilde{g}_1^\tau = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \langle \beta_{1,3}^*, \tilde{g}_1^{\mathcal{T}} \rangle \beta_{1,1}^* \circ \beta_{1,2}^* = \gamma^{*\tau} \beta_{1,1}^* \circ \beta_{1,2}^* = \gamma^{*\tau} \beta_{1,1}^* \beta_{1,2}^{*T}. \quad \text{(B.17)}$$

Therefore, $\gamma^{*\mathcal{T}}$ is the singular vector of $\mathcal{T}^* \times_3 \tilde{g}_1^\tau$ when rank $r = 1$.
Next, we need to prove that there exists some τ such that $\gamma^{*\tau}$ is sufficiently separated from 0. Since $\tilde{g}_1^\tau = U_1 U_1^T g_1^\tau$, we have

$$\gamma^{*\tau} = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \langle U_1 U_1^T \beta_{1,3}^*, g_1^\tau \rangle,$$

where U_1 is defined as the rank-1 eigen-decomposition of B_1, and $B_1 = \Pi_{\text{off-diag}}(A_1 A_1^T)$.

Since g_1^τ is a standard Gaussian vector,

$$\mathbb{E}(\gamma^{*\tau} | \Omega_1, \mathcal{E}_1, \ldots, \Omega_n, \mathcal{E}_n) = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \langle U_1 U_1^T \beta_{1,3}^*, \mathbb{E}(g_1^\tau) \rangle = 0$$

$$\mathbb{V}\text{ar}(\gamma^{*\tau} | \Omega_1, \mathcal{E}_1, \ldots, \Omega_n, \mathcal{E}_n) = w_1^{*2} \left(\sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right)^2 \left\| U_1 U_1^T \beta_{1,3}^* \right\|^2 \mathbb{V}ar(g_1^\tau | \Omega_i, \mathcal{E}_i)).$$

Then

$$\mathbb{V}\text{ar}(\gamma^{*\tau} | \Omega_1, \mathcal{E}_1, \ldots, \Omega_n, \mathcal{E}_n) = w_1^{*2} \left(\sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right)^2 \left\| U_1 U_1^T \beta_{1,3}^* \right\|^2.$$

Without loss of generality, let $\gamma^{*1} \geq \gamma^{*2} \geq \cdots \geq \gamma^{*L}$. By Lemma A.8 and $r = 1$, for any fixed small constant $\delta > 0$, with probability at least $1 - \delta$, we have

$$\gamma^{*1} \gtrsim w_1^* \left| \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right| \| U_1 U_1^T \beta_{1,3}^* \|, \quad \text{(B.18)}$$

which holds due to the condition that $L \geq C_1'$ for some constant C_1'. Let $A_1^* = \text{unfold}_3(\mathcal{T}^*)$, then

$$A_1^* = \text{unfold}_3(w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^*).$$

Then

$$A_1^* = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \beta_{1,3}^* (\beta_{1,1}^* \otimes \beta_{1,2}^*)^T \in \mathbb{R}^{d_3 \times d_1 d_2}.$$

Let U_1^* be the basis of the column space of $A_1^* A_1^{*T}$. Intuitively, the space spanned by U_1 is close to the space spanned by the true tensor factor. Then, by applying Lemma A.4 with $\delta = \|U_1 - U_1^*\|$, $V = U_1, U = U_1^*$, $u_0 = \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \beta_{1,3}^*$ the bound in (B.18) can be simplified as,

$$\gamma^{*1} \gtrsim w_1^* \sqrt{1 - \|U_1 - U_1^*\|^2}, \tag{B.19}$$

From the proof of Theorem 1 in Cai et al, 2021 [18] we have,

$$\|U_1 - U_1^*\| \leq \sqrt{2} \|U_1 U_1^T - U_1^* U_1^{*T}\| \leq \frac{\|A_1^* A_1^{*T} - A_1 A_1^T\|}{\sigma(A_1^*)^2}, \tag{B.20}$$

where $\sigma(A_1^*)$ is the singular value of A_1^*. Since rank $r = 1$, A_1^* is a rank-1 matrix. Then, we have $\sigma(A_1^*) = w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i$. Also, using the results from Cai et al, 2021 [18], we have

$$\begin{aligned}
\|A_1^* A_1^{*T} - A_1 A_1^T\| \lesssim & \left\{ \frac{\|A_1^*\|_{2,\infty} + \sigma\sqrt{\tilde{d}}}{\sqrt{p}} + \frac{\|A_1^{*T}\|_{2,\infty} + \sigma\sqrt{\tilde{d}}}{\sqrt{p}} \right\} \times \\
& \left\{ \frac{\|A_1^{*T}\|_{2,\infty} + \sigma\sqrt{\tilde{d}}}{\sqrt{p}} + \|A_1^{*T}\|_{2,\infty} \right\} \log(\tilde{d}) \\
& + \frac{\|A_1^{*T}\|_{2,\infty} + \sigma\sqrt{\tilde{d}}}{\sqrt{p}} \sqrt{\log(\tilde{d})} \|A_1^*\| + \|A_1^*\|_{2,\infty},
\end{aligned} \tag{B.21}$$

where $\|A\|_{2,\infty} = \max_{i \in [m]} \|A_{i,:}\|_2$ for any matrix $A \in \mathbb{R}^{m \times n}$ and $\tilde{d} = \max\{d_3, d_1 d_2\}$.

$$\begin{aligned}
\|A_1^*\|_{2,\infty} &= \max_{l \in [d_3]} \left\| \beta_{1,3,l}^* w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i (\beta_{1,1}^* \otimes \beta_{1,2}^*)^T \right\|_2 \\
&\leq \max_{l \in [d_3]} |\beta_{1,3,l}^*| \left| w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right| \left\| (\beta_{1,1}^* \otimes \beta_{1,2}^*)^T \right\|_2 \leq c_1 w_1^* \frac{\mu}{\sqrt{s}},
\end{aligned} \tag{B.22}$$

using Assumption 1: $\max|\beta_{1,3,l}^*| \leq \frac{\mu}{\sqrt{s}}$, $\left| w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right| \leq c_1 w_1^*$, $\left\| (\beta_{1,1}^* \otimes \beta_{1,2}^*)^T \right\|_2 \leq 1$.

In addition, we have

$$\|A_1^{*T}\|_{2,\infty} \leq \left| w_1^* \sum_{i=1}^{n} \frac{1}{n} \beta_{1,4}^{*T} x_i \right| \|\beta_{1,3}^{*T}\|_2 \max_{l_1 \in [d_1], l_2 \in [d_2]} |\beta_{1,1,l_1}^* \beta_{1,2,l_2}^*|.$$

Then,

$$\|A_1^{*T}\|_{2,\infty} \leq c_1 w_1^* \max_{l_1 \in [d_1], l_2 \in [d_2]} |\beta^*_{1,1,l_1} \beta^*_{1,2,l_2}| \leq c_1 w_1^* \frac{\mu}{\sqrt{s}}. \tag{B.23}$$

Recall that A_1^* is a rank-1 matrix, then, we have

$$\|A_1^*\|_2 = \|A_1^*\|_F = \left| w_1^* \sum_{i=1}^n \frac{1}{n} \beta_{1,4}^{*T} x_i \right| \|\beta^*_{1,3}\| \left\| (\beta^*_{1,1} \otimes \beta^*_{1,2})^T \right\| \leq c_1 w_1^*. \tag{B.24}$$

Combining (B.21) - (B.24) and simplifying the formula, we have

$$\|A_1^* A_1^{*T} - A_1 A_1^T\| \lesssim \frac{w_1^{*2}\mu^3}{s^{1.5}p} \log(\tilde{d}) + \frac{\sigma^2 \tilde{d} \log \tilde{d}}{p} + \frac{w_1^{*2}\mu^2}{s} \sqrt{\frac{\log(\tilde{d})}{p}} + w_1^* \sigma \sqrt{\frac{\tilde{d}\log(\tilde{d})}{p}} + \frac{w_1^{*2}\mu^2}{s}.$$

Therefore, for any arbitrary small constant $\delta > 0$, with probability greater than $1 - \delta$,

$$\gamma^{*1} \gtrsim w_1^*. \tag{B.25}$$

We have shown that there exists some $\tau \in [L]$, such that $\gamma^{*\tau} \gtrsim w_1^*$. This means that $\beta^*_{1,3}$ exhibits the largest correlation with the projected g_1, which further implies that $\beta_{1,1}$ is the largest left singular vector of $\mathcal{T}^* \times_3 \tilde{g}_1^\tau$.

Step 2:

In Step 2, we prove that the top singular vector v_1^τ is close to $\beta^*_{1,1}$. We obtain the bound for v_1 as an example, while the bound for v_2 and v_3 can be derived similarly.

Recall that v_1^τ is the top left singular vector of M^τ, where

$$M^\tau = \frac{1}{p}\mathcal{T} \times \tilde{g}_1^\tau = \frac{1}{p}\mathcal{T}^* \times \tilde{g}_1^\tau + \frac{1}{p}\{\mathcal{T} - \mathcal{T}^*\} \times \tilde{g}_1^\tau = \frac{1}{p} \underbrace{\gamma^{*\tau} \beta^*_{1,1} \beta^{*T}_{1,2}}_{M^{*\tau}} + \frac{1}{p}\{\mathcal{T} - \mathcal{T}^*\} \times \tilde{g}_1^\tau,$$

where the second inequality is due to the definition of $\gamma^{*\tau}$ in (B.17).

Using Theorem A.4 (Wedin's theorem), we get

$$\|v_1^\tau - \beta^*_{1,1}\| \leq \frac{\|(M^\tau - M^{*\tau})\beta^*_{1,1}\|_2}{\gamma^{*\tau} - \|M^\tau - M^{*\tau}\|}. \tag{B.26}$$

Next, we bound $\|(M^\tau - M^{*\tau})\beta^*_{1,1}\|_2$ and $\|M^\tau - M^{*\tau}\|$.

First, recall that $\tilde{g}_1^\tau = U_1 U_1^T g_1^\tau$. Define $\tilde{g}_1^{*\tau} = U_1^* U_1^{*T} g_1^{*\tau}$, and decompose,

$$M^\tau - M^{*\tau} = \{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_3 \tilde{g}_1^\tau = \underbrace{\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_3 g_1^{*\tau}}_{V_1} + \underbrace{\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_3 \{\tilde{g}_1^\tau - g_1^{*\tau}\}}_{V_2}.$$

Note that V_1 is a zero mean random matrix in $\mathbb{R}^{d_1 \times d_2}$

$$V_{1,l_1,l_2} = \sum_{l_3 \in [d_3], i \in [n]} \frac{1}{n} g^{*\tau}_{1,l_3} [\{p^{-1}\delta_{i,l_1,l_2,l_3} - 1\} \beta^{*T}_{1,4} x_i w_1^* \beta^*_{1,1,l_1} \beta^*_{1,2,l_2} \beta^*_{1,3,l_3} + p^{-1} \mathcal{E}_{i,l_1,l_2,l_3}].$$

By Lemma A.9, with probability $1-\mathcal{O}(d^{-10})$, we have

$$\|V_1\| \lesssim \frac{\|g_1^{*\tau}\|_\infty \sigma \log^{2.5}(d)}{pn} + \frac{\|g_1^{*\tau}\|_\infty w_1^* \mu \sqrt{\log(d)}}{\sqrt{pns}} + \frac{\|g_1^{*\tau}\|_2 \sigma \sqrt{s\log(d)}}{\sqrt{pn}}.$$

Further, with probability at least $1-\mathcal{O}(d^{-20})$, $\|g_1^{*\tau}\|_\infty = \|U_1^* U_1^{*T} g_1\|_\infty$ $\lesssim \|U_1^*\|_{2,\infty}\sqrt{\log(d)} \lesssim \mu\sqrt{\log(d)/s}$, and $\|g_1^{*\tau}\|_2 \lesssim \|U_1^*\|_F \sqrt{\log(d)} \lesssim \sqrt{\log(d)}$. Then, by applying Assumption 2(d), we get

$$\|V_1\| \lesssim \frac{\mu\sigma\log^3(d)}{pn\sqrt{s}} + \frac{w_1^*\mu^2\log(d)}{\sqrt{pns^2}} + \frac{\sigma\sqrt{s}\log(d)}{\sqrt{pn}} \lesssim \frac{w_1^*\mu^2\log(d)}{\sqrt{pns^2}} + \frac{\sigma\sqrt{s}\log(d)}{\sqrt{pn}}. \quad \text{(B.27)}$$

Next, we turn to V_2, and have that

$$\|V_2\| \leq \|\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_3 \{\tilde{g}_1^\tau - g_1^{*\tau}\}\| \leq \|p^{-1}\mathcal{T} - \mathcal{T}^*\| \|\tilde{g}_1^\tau - g_1^{*\tau}\|_2.$$

By applying Proposition A.19 and Lemma A.11, we have that with probability at least $1-\mathcal{O}(d^{-12})$,

$$\|\tilde{g}_1^\tau - g_1^{*\tau}\|_2 \lesssim \|(U_1U_1^T - U_1^*U_1^{*T})g^\tau\| \lesssim \|U_1U_1^T - U_1^*U_1^{*T}\|\sqrt{\log(d)} \ll 1.$$

Moreover,

$$\|p^{-1}\mathcal{T} - \mathcal{T}^*\| \leq \left\| \frac{w_1^*}{pn} \sum_i \Pi_{\Omega_i}\left(\beta_{1,4}^{*T} x_i \beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^*\right) - \mathcal{T}^* \right\| + \left\| \frac{1}{pn}\sum_i \Pi_{\Omega_i}(\mathcal{E}_i) \right\|. \quad \text{(B.28)}$$

By Lemma A.10, with probability at least $1-\mathcal{O}(d^{-10})$, we have

$$\left\| \frac{1}{pn}\sum_i \Pi_{\Omega_i}(\mathcal{E}_i) \right\| \leq \frac{1}{pn}\sum_i \|\Pi_{\Omega_i}(\mathcal{E}_i)\| \lesssim \frac{\sigma\log^{7/2}(d)}{p} + \sigma\sqrt{\frac{d\log^5(d)}{p}}. \quad \text{(B.29)}$$

By the same Lemma A.10, with probability at least $1-\mathcal{O}(d^{-10})$, we have

$$\left\| \frac{w_1^*}{pn} \sum_i \Pi_{\Omega_i}\left(\beta_{1,4}^{*T} x_i \beta_{1,1}^* \circ \beta_{1,2}^* \circ \beta_{1,3}^*\right) - \mathcal{T}^* \right\| \lesssim \frac{w_1^*\mu^3\log^3(d)}{s^{1.5}p} + \frac{w_1^*\mu^2\log^{5/2}(d)}{\sqrt{pn}s}. \quad \text{(B.30)}$$

Combining (B.28) - (B.30), we have

$$\|V_2\| \lesssim \|\|p^{-1}\mathcal{T} - \mathcal{T}^*\| \lesssim \frac{w_1^*\mu^3\log^3(d)}{s^{1.5}p} + \frac{w_1^*\mu^2\log^{5/2}(d)}{\sqrt{pn}s} + \frac{\sigma\log^{7/2}(d)}{p} + \sigma\sqrt{\frac{d\log^5(d)}{p}}.$$

Combining the bounds of V_1 and V_2, we obtain that

$$\|M^\tau - M^{*\tau}\| \leq \|V_1\| + \|V_2\| \ll w_1^*.$$

Second, to bound $\|(M^\tau - M^{*\tau})\beta^*_{1,1}\|_2$, we have, by the definition of the operator norm,

$$\|(M^\tau - M^{*\tau})\beta^*_{1,1}\|_2 \leq \|\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_1 \beta^*_{1,1} \times_3 \tilde{g}^\tau_1\| \leq \|\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_1 \beta^*_{1,1}\|\|\tilde{g}^\tau_1\|.$$

By Lemma A.9, we have

$$\|\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_1 \beta^*_{1,1}\| \lesssim \frac{\|\beta^*_{1,1}\|_\infty \sigma \log^{2.5}(d)}{pn} + \frac{\|\beta^*_{1,1}|_\infty w^*_1 \mu\sqrt{\log(d)}}{\sqrt{pns}} + \frac{\|\beta^*_{1,1}\|_2 \sigma\sqrt{s\log(d)}}{\sqrt{pn}}.$$

Then, by Assumptions 1(c) and 2(d), we have

$$\begin{aligned}\|\{p^{-1}\mathcal{T} - \mathcal{T}^*\} \times_1 \beta^*_{1,1}\| &\lesssim \frac{\mu\sigma\log^{2.5}(d)}{pn\sqrt{s}} + \frac{w^*_1\mu^2\log(d)}{\sqrt{pns^2}} + \frac{\sigma\sqrt{s}\log(d)}{\sqrt{pn}} \\ &\lesssim \frac{w^*_1\mu^2\log(d)}{\sqrt{pns^2}} + \frac{\sigma\sqrt{s}\log(d)}{\sqrt{pn}}.\end{aligned}$$

We have proved in (B.25) that $\gamma^{*\tau} \gtrsim w^*_1$. Besides, we have $\|M^\tau - M^{*\tau}\| \ll w^*_1$. Therefore, the difference in (B.26) becomes,

$$\|v^\tau_1 - \beta^*_{1,1}\| \lesssim \mu^2\sqrt{\frac{\log(d)}{pns^2}} + \frac{\sigma}{w^*_1}\sqrt{\frac{s\log(d)}{pn}}. \tag{B.31}$$

Next, we show that v^τ_1 is $c\mu$-mass vector, where c is a general constant.

$$\max_{l_1\in[d_1]}|v^\tau_{1,l_1}| \leq \max_{l_1\in[d_1]}|\beta^*_{1,l_1}| + c'\mu^2\sqrt{\frac{\log(d)}{pns^2}} + c''\frac{\sigma}{w^*_1}\sqrt{\frac{s\log(d)}{pn}} \leq \frac{\mu}{\sqrt{s}} + \frac{c'_1\mu}{\sqrt{s}} + \frac{c''_1\mu}{\sqrt{s}},$$

where the first inequality holds by (B.31). The second inequality holds due to assumption on the sample size.

Step 3:

In Step 3, we prove that $\hat{\beta}^{(0)}_{1,4}$ is close to true factor.

We have shown in (B.10) that, if $p \gtrsim \mu^3\log(d)/\{ns^{1.5}\}$, with a high probability,

$$\|\hat{\beta}_{1,4} - \beta^*_{1,4}\| \leq k\epsilon + \frac{\tilde{C}_2\sigma}{\lambda_{\min}w^*_1}\sqrt{\frac{qs\log(d)}{np}},$$

where k is some constant, and ϵ is the estimator error of v_j in (B.31). Therefore, the final error is

$$\max\{\max_j\|\hat{\beta}^{(0)}_{1,j} - \beta^*_{1,j}\|_2\} \lesssim \mu^2\sqrt{\frac{\log(d)}{nps^2}} + \frac{\sigma}{w_1*}\sqrt{\frac{s\log(d)}{np}}.$$

That completes the proof of Theorem 5.3.

□

Vita Auctoris

Dinara Miftyakhetdinova was born in 1999 in Moscow, Russia. She graduated from Lyceum 1571 in 2016. From there she went on to the Lomonosov Moscow State University where she obtained a Bachelor's degree in Applied Mathematics and Computed Science in 2020. She is currently a candidate for the Master's degree in Mathematics and Statistics at the University of Windsor and hopes to graduate in Fall 2022.