Causal Customer Churn Analysis with Low-rank Tensor Block Hazard Model

Chenyin Gao [1] **Zhiming Zhang** [2] **Shu Yang** [1]

Abstract

This study introduces an innovative method for analyzing the impact of various interventions on customer churn, using the potential outcomes framework. We present a new causal model, the tensorized latent factor block hazard model, which incorporates tensor completion methods for a principled causal analysis of customer churn. A crucial element of our approach is the formulation of a 1-bit tensor completion for the parameter tensor. This captures hidden customer characteristics and temporal elements from churn records, effectively addressing the binary nature of churn data and its time-monotonic trends. Our model also uniquely categorizes interventions by their similar impacts, enhancing the precision and practicality of implementing customer retention strategies. For computational efficiency, we apply a projected gradient descent algorithm combined with spectral clustering. We lay down the theoretical groundwork for our model, including its non-asymptotic properties. The efficacy and superiority of our model are further validated through comprehensive experiments on both simulated and real-world applications.

1. Introduction

Customer retention, loyalty, and churn are increasingly important topics across industries. Customer churn, or attrition, occurs when customers disengage from a company by ending subscriptions, moving to competitors, or changing their buying habits (Buckinx et al., 2007). Analyzing and preventing churn is critical for sustainable growth and profitability. Companies often explore diverse retention strategies, such as customized incentives, to extend customer lifetimes. Understanding the causality behind customer churn helps companies adopt proper retention strategies and essentially increase the customer lifetime value and the company's business value.

In our approach to analyzing causal churn, we define potential outcomes as the indicators of churn over time under different levels of treatment. Thus, the potential outcomes can be envisioned as a three-dimensional tensor, indexed by the customer, time, and intervention, where each entry represents whether a customer would churn at a particular time under a specific treatment (see Figure 1). In practice, only actual churn trajectories are observed. To address this, tensor completion methods can be employed to fill in all missing potential outcomes and facilitate the analysis of various treatment strategies. Tensors effectively uncover the hidden multiway data structure, often using low-rankness, which decomposes the tensor into a low-dimensional core tensor and matrix factors for each dimension. However, applying these existing tensor completion methods presents several challenges in our context.

Firstly, unlike tensors with continuous values, our churn analysis tensor is binary, with entries as $0/1$ indicators. Moreover, once a customer churns at a certain time, it remains churned, resulting in a monotone churn pattern. Therefore, applying low-rankness to the original potential outcome tensor may not be suitable in this case. Secondly, many retention interventions might have similar effects on customers, suggesting a potential benefit in grouping these interventions for more accurate estimation and streamlined implementation in practice. Previous research has proposed integrating interventions based on prior knowledge (Laber et al., 2014; Liu et al., 2018; Pan & Zhao, 2021). However, a data-driven approach to identify and cluster interventions with similar effects is desirable. For instance, Ma et al. (2022) introduced a method using adaptive fusion penalty for clustering interventions, but it is limited to single-stage outcomes and parametric treatment effect models. Exploring and autonomously grouping interventions with similar effects over time to streamline the dimension space is therefore a significant area of interest. Finally, a major challenge arises from the uniform missing mechanisms in current matrix/tensor completion methods, which fail to consider the endogeneity in treatment assignment. As a result, directly applying these methods could lead to confounding biases.

[1]Department of Statistics, North Carolina State University, Raleigh, U.S.A. [2]Independent scholar, Iowa State University, Ames, U.S.A.. Correspondence to: Shu Yang <syang24@ncsu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

We develop a novel tensorized latent factor block hazard model in (3) to conduct causal analysis on customer churn trajectories in relation to various treatment strategies. This model redefines the problem into a 1-bit tensor completion problem for the parameter tensor, leveraging structural information, such as low rankness and clustering blocks. In particular, it captures the customer attributes and temporal features by the latent factors inferred from the churn history. Besides, a block structure is adopted for the interventions based on their impact on the churn statuses to reduce the number of interventions. This data-driven clustering allows us to automatically identify the homogeneous intervention groups. To our knowledge, little work has explored the causality of customer churn analysis with grouped latent factors and we are among the first to provide non-asymptotic error analysis for this low-rank tensor block hazard model.

Our contributions are summarized as follows:

1) Tensorized Latent Factor Block Hazard Model: We introduce a model applying low-rankness to the parameter tensor (rather than the data tensor) to leverage latent unit and temporal factors and identify homogenous intervention groups more effectively.

2) Computational Methodology: We use a projected gradient descent algorithm with spectral clustering to solve the inverse probability treatment weighted (IPTW) loss, adjusting for confounding effects and scaling well to large datasets.

3) Optimal Treatment Search and Learning: The proposed framework provide the survival probabilities under all interventions and thus enables one to identify the individual optimal interventions that maximize retention time, closely related to the optimal policy search and Q-learning in causal inference (Qian & Murphy, 2011); and

4) Theoretical Underpinnings and Empirical Evidence: We have established the non-asymptotic properties of our proposed model, including the upper bound on the tensor recovery accuracy and the clustering misclassification rate. Besides, we demonstrate the practical benefits and effectiveness of our framework via comprehensive synthetic experiments and a real-data application. Our implementation codes will be made publicly available after the acceptance of this manuscript.

2. Related Work

Customer churn prediction can be naturally perceived as a classification task. With the rise of machine learning algorithms, various methods have been proposed for churn analysis, including support vector machines (Coussement & Van den Poel, 2008), random forest (Xie et al., 2009), and other ensemble methods such as bagging and boosting (Lu et al., 2012). Deep neural networks (DNN) have also been employed to extract valuable features related to customer churn, which significantly improve the performance on real-world datasets (Mishra & Reddy, 2017; Zhang et al., 2017; Umayaparvathi & Iyakutti, 2017; Rudd et al., 2021). However, these classification models mainly focus on determining churn status at a fixed time point, often overlooking the information contained in time to churn.

Considering customer lifetime as a time-to-churn metric and using survival analysis offers deeper insights into customer-company engagement (Lu, 2002; Larivière & Van den Poel, 2004). Traditional survival analysis, though, relies on potentially restrictive assumptions about hazard functions. Advanced survival models like survival random forests (Ishwaran et al., 2008), Cox boosting (Binder et al., 2009), survival Super-Learner (Van der Laan & Rose, 2011), and time-to-event reinforcement learning (Maystre & Russo, 2022) have been developed to better handle more complex data. Similarly, DNNs have also been utilized for survival analysis to handle the special loss function induced by the censored data (Zhu et al., 2016; Katzman et al., 2018; Ching et al., 2018; Zhao & Feng, 2020). However, these models may not accurately reflect the causal impact of interventions on churn due to confounding biases (Yang, 2021). For example, if incentives are offered only to at-risk customers, it may be falsely ineffective as this group naturally shows higher retention and shorter churn times compared to others. Therefore, a more principled approach is necessary for the causal analysis of retention interventions.

We use the potential outcomes framework of different intervention strategies for churn analysis. Here, the potential outcome is defined as the potential outcome (possibly contrary to fact) had the unit (customer) received a specific treatment (intervention). The fundamental problem of causal inference is that each unit receives only one treatment, leaving other potential outcomes unknown. Thus, the causal analysis problem is essentially a missing data problem. To address this, matrix or tensor completion techniques (Davenport et al., 2014; Mao et al., 2023), commonly used for filling in missing data, are applicable. Tensor completion problem is initially addressed by unfolding tensors into matrices (Tomioka et al., 2010; Gandy et al., 2011; Liu et al., 2012). However, such unfolding-based methods might discard the multi-way structure of the tensor, rendering them less efficient. A non-convex approach is motivated to solve this problem in Xia & Yuan (2017); Xia et al. (2021); Cai et al. (2021), where the authors apply low-rank tensor factorization to enforce the low-rankness and update the factors iteratively.

Most of these matrix or tensor completion methods assume that the missingness occurs completely at random. However,

the treatment assignment is typically an endogenous process, as it may be affected by some prognostic factors of that unit (Pearl, 2009). A stream of work employing the debiased matrix/tensor completion method has been proposed in Mandal & Parkes (2019); Agarwal et al. (2020); Athey et al. (2021); Agarwal et al. (2021); Mao et al. (2023), which re-weights each unit inversely to its probability of being assigned the actualized treatments. Admittedly, it is more challenging to recover the low-rank latent parameter matrix/tensor from the binary outcomes. A few methods have been proposed to handle such quantized and possibly corrupted outcomes, utilizing the properties of the matrix/tensor norm (Davenport et al., 2014; Cai & Zhou, 2013) or enforcing the latent parameters lying in a low-rank matrix/tensor subspace (Wang & Li, 2020; Ashraphijuo & Wang, 2020; Mao et al., 2024).

3. Basic setup

3.1. Notation and preliminaries

Before presenting our framework, let us present some basics of tensor algebra. Scalars are denoted by lowercase letters (e.g., x, y), while vectors and matrices use bold lowercase (e.g., $\boldsymbol{x}, \boldsymbol{y}$) and uppercase letters (e.g., $\boldsymbol{X}, \boldsymbol{Y}$), respectively. The outer product of vectors $\boldsymbol{x} \in \mathbb{R}^{p_1}$ and $\boldsymbol{y} \in \mathbb{R}^{p_2}$ is $\boldsymbol{x} \otimes \boldsymbol{y} \in \mathbb{R}^{p_1 \times p_2}$. Higher-order tensors are represented by calligraphic letters (e.g., $\mathcal{X}, \mathcal{Y}$). The entry-wise and inner products of tensors $\mathcal{X}$ and $\mathcal{Y}$ are $\mathcal{X} \odot \mathcal{Y}$ and $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1,i_2,i_3} \mathcal{X}_{i_1,i_2,i_3} \mathcal{Y}_{i_1,i_2,i_3}$, respectively.

Tensors are unfolded into matrices on mode-k using operator $\mathcal{M}_{(k)}(\cdot)$. For example, after unfolding the tensor $\mathcal{X}$ on its first mode, we have $\mathcal{M}_{(1)}(\mathcal{X}) \in \mathbb{R}^{p_1 \times p_2 p_3}$, where $[\mathcal{M}_{(1)}(\mathcal{X})]_{i_1, i_2 + p_2(i_3 - 1)} = \mathcal{X}_{i_1,i_2,i_3}$. Also, mode-$k$ Tensor-matrix products are denoted by $\mathcal{X} \times_k \boldsymbol{U}_k$. For example, let $\boldsymbol{U}_1 \in \mathbb{R}^{r_1 \times p_1}$, the mode-1 tensor-matrix multiplication is $\mathcal{X} \times_1 \boldsymbol{U}_1 \in \mathbb{R}^{r_1 \times p_2 \times p_3}$, where $(\mathcal{X} \times_1 \boldsymbol{U}_1)_{i_1,i_2,i_3} = \sum_{j_1=1}^{p_1} \mathcal{X}_{j_1,i_2,i_3} \boldsymbol{U}_{i_1,j_1}$. The multi-linear rank of a tensor $\mathcal{X}$ is $(r_1, r_2, r_3) = \{\text{rank}(\mathcal{M}_{(1)}(\mathcal{X})), \text{rank}(\mathcal{M}_{(2)}(\mathcal{X})), \text{rank}(\mathcal{M}_{(3)}(\mathcal{X}))\}$. Tucker decomposition (Tucker, 1966) factorizes a tensor $\mathcal{X}$ into a core tensor $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and orthogonal matrices $\boldsymbol{U}_i \in \mathbb{R}^{p_i \times r_i}$ as

$$\mathcal{X} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 = [\![\mathcal{S}; \boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3]\!]. \quad (1)$$

In this decomposition, $\mathcal{S}$ can be considered as the principal components with $\boldsymbol{U}_i$ being the mode-i loading matrices. Tensor norms including spectral norm $\|\mathcal{X}\|$, nuclear norm $\|\mathcal{X}\|_*$, Frobenius norm $\|\mathcal{X}\|_F$, and max norm $\|\mathcal{X}\|_{\max}$ are used; see Kolda & Bader (2009) for a comprehensive review. Finally, c_0, C_0 represent generic positive constants, $\asymp$ ($\lesssim$ and $\gtrsim$) indicates equality (inequality) up to multiplicative numerical constants, and $[r]$ denotes the r-set $\{1, \cdots, r\}$.

3.2. Potential outcomes and causal assumptions

For each unit i, $\mathbf{V}_i = (\boldsymbol{X}_i, \boldsymbol{A}_i, \delta_i, \boldsymbol{Y}_i)$ represent a d-dimensional covariate vector, a k-dimensional binary treatment vector $\boldsymbol{A}_i = (A_{i,1}, \cdots, A_{i,k})$, a churn indicator δ_i taking values from $\{0, 1\}$, and a T-dimensional vector of retention statutes with possible censoring over T time points $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,T})^{\mathrm{T}}$, respectively. The censorship of churn statuses is determined by the churn indicator δ_i. For examples, $(\boldsymbol{Y}_i, \delta_i) = \{(1, 1, 0, \ldots, 0)^{\mathrm{T}}, 1\}$ implies that customer i is retained until time point 2 and churns at time point 3, i.e., the churn status is observed; $(\boldsymbol{Y}_i, \delta_i) = \{(1, 1, 0, \ldots, 0)^{\mathrm{T}}, 0\}$ implies that customer i stays for the first two time points and does not churn, i.e., the churn status is censored.

Under the potential outcomes framework (Rubin, 1974), $\boldsymbol{Y}_i^{(\boldsymbol{a})}$ denotes the potential churn trajectory of unit i had the treatment be set to $\boldsymbol{a} \in \mathbb{A}$, where $\mathbb{A}$ contains all possible binary treatment vectors of dimension k, i.e., a exhaustive set of size 2^k. The "survival" function for retention at each time point t is defined as $S_i^{(\boldsymbol{a})}(t) = \mathbb{E}(Y_{i,t}^{(\boldsymbol{a})}) = \mathbb{P}(Y_{i,t}^{(\boldsymbol{a})} = 1)$ under treatment $\boldsymbol{a}$. Similarly, the treatment-specific lifetime is defined by $\sum_{t=1}^{T} Y_{i,t}^{(\boldsymbol{a})}$. Since each unit receive only one treatment, not all potential outcomes are observable, necessitating certain assumptions for causal analysis:

A1) (Stable Unit Treatment Value) $\boldsymbol{Y}_i = \boldsymbol{Y}_i^{(\boldsymbol{a})}$ for $\boldsymbol{A}_i = \boldsymbol{a}$;

A2) (No Unmeasured Confounders) $\boldsymbol{A}_i \perp\!\!\!\perp \boldsymbol{Y}_i^{(\boldsymbol{a})} \mid \boldsymbol{X}_i$ for all $\boldsymbol{a} \in \mathbb{A}$

A3) (Non-informative Censoring) $\delta_i \perp\!\!\!\perp \boldsymbol{Y}_i^{(\boldsymbol{a})} \mid \boldsymbol{X}_i, \boldsymbol{A}_i$; and

A4) (Positivity) The generalized propensity score $\pi(\boldsymbol{a} \mid \boldsymbol{X}_i) = \mathbb{P}(\boldsymbol{A}_i = \boldsymbol{a} \mid \boldsymbol{X}_i)$ for any $\boldsymbol{a} \in \mathbb{A}$ is bounded away from 0 and 1 almost surely.

A1) rules out interference between units and multiple versions of treatment, A2) requires that $\boldsymbol{X}_i$ accounts for all variables influencing both the treatment uptake and outcome, A3) is a common censoring at random assumption for survival analysis, which is a special case of the coarsening at random (Tsiatis, 2006); and A4) ensures that every unit has a non-zero probability of receiving each level of treatment. Under Assumptions A1)–A4), the survival probability $S^{(\boldsymbol{a})}(t)$ are identifiable under all interventions. Moreover, in discrete survival analysis, we approximate the expected treatment-specific lifetime by the sum of the survival probabilities up to T, that is, $\sum_{t=1}^{T} S^{(\boldsymbol{a})}(t)$. Therefore, the individual optimal treatment D_{opt} can be derived from $D_{i,\text{opt}} = \arg\max_{\boldsymbol{a} \in \mathbb{A}} \sum_{t=1}^{T} S_i^{(\boldsymbol{a})}(t)$ for the i-th customer.

3.3. Tensorized hazard model

We model potential churn statuses of N units over T time points and L treatments as a 3-mode tensor

$$\mathcal{Y} = \left(Y_{i,t}^{(\boldsymbol{a})}\right)_{i=1,\cdots,N,\quad t=1,\cdots,T,\quad \boldsymbol{a}\in\mathbb{A}} = (\mathcal{Y}_{i,t,l}),$$

where $\boldsymbol{a}$ is a binary treatment vector converted to decimal form l, e.g., $l = (00011)_{10} = 3$ for $\boldsymbol{a} = (00011)$. Likewise, $(l)_2 = \boldsymbol{a}$ is for binary conversion. Since $\boldsymbol{a}$ and l have one-to-one correspondence, we will use them interchangeably.

Our objective is to leverage the tensor structures to impute the missing potential outcomes in $\mathcal{Y}$ and provide valid estimates for the causal parameters. As the entries of the potential outcome tensor, representing churn status, are binary (0 or 1) and demonstrate a time-monotone pattern, direct low-rank constraints on $\mathcal{Y}$ are inappropriate. We propose a low-rank hazard model for $\mathcal{Y} \mid \Theta \sim \text{Bernoulli}\{\mathbb{P}(\mathcal{Y} = 1 \mid \Theta)\}$, where $\Theta = (\theta_{i,t,l})$ is an unknown parameter tensor of the same dimension as $\mathcal{Y}$. In this model, $\theta_{i,t,l}$ is the parameter in the hazard probability

$$\mathbb{P}\left(\mathcal{Y}_{i,t,l} = 1 \mid \mathcal{Y}_{i,t-1,l} = 1, \boldsymbol{X}_i\right) = f(\theta_{i,t,l}), \qquad (2)$$

with $f(\cdot)$ being a strictly increasing and log-concave link function. We further assume that the link function $f(\cdot)$ is twice differentiable, satisfying $f(\theta) + f(-\theta) = 1$, $f'(\theta) = f'(-\theta)$ and $f''(\theta) = -f''(-\theta)$. These conditions are typically employed in the community of one-bit matrix/tensor completion and some common choices of $f(\cdot)$ that satisfy these conditions include the logistic link, the probit link, and the Laplacian link (Wang & Li, 2020). The individual probability of retention at one time point t is $S_i^{(\boldsymbol{a})}(t) = \prod_{s=1}^{t} f(\theta_{i,s,l})$ for treatment $a = (l)_2$, ensuring a decreasing retention probability over time.

Although the individual survival probabilities $S_i^{(\boldsymbol{a})}(t)$ can be approximated, we only observe $\mathcal{Y}_{i,t,l}$, which is a binary and quantized version of $\theta_{i,t,l}$. This is similar to the threshold model used in 1-bit matrix/tensor completion (Cai & Zhou, 2013; Ghadermarzy et al., 2018); see Figure 1. As Θ is latent, structural assumptions are needed for its identification and estimation.

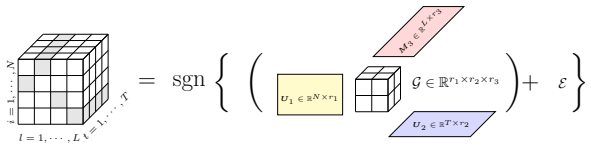

Figure 1. A new tensor representation of potential outcomes with three modes (customer $\times$ time $\times$ intervention).

3.4. Low-rank structure and treatment clustering

Tensor structures, particularly low-rankness, aid in parameter identification. Additionally, in a large treatment space, certain treatments may exhibit similar effects. Often, retention strategies involve a mix of various incentives, and altering one or two of these incentives might not significantly impact the overall effectiveness of the strategy. Additionally, some retention strategies, though targeting different aspects, are based on similar behavioral models and mechanisms. Identifying and clustering these treatments can reduce the treatment space.

We assume our parameter tensor, Θ, admits the latent factor block model with a single discrete structure on the third mode:

$$\begin{aligned}\Theta &= \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{M} \\ &= \sum_{j_1=1}^{r_1}\sum_{j_2=1}^{r_2}\sum_{j_3=1}^{r_3} \mathcal{S}_{j_1 j_2 j_3} \boldsymbol{u}_{1,j_1} \otimes \boldsymbol{u}_{2,j_2} \otimes \boldsymbol{m}_{j_3}, \qquad (3)\end{aligned}$$

with $\mathcal{S} \in \mathbb{R}^{r_1\times r_2\times r_3}$ as the core tensor, $\boldsymbol{U}_1 = (\cdots \boldsymbol{u}_{1,j_1} \cdots) \in \mathbb{R}^{N\times r_1}$ and $\boldsymbol{U}_2 = (\cdots \boldsymbol{u}_{2,j_2} \cdots) \in \mathbb{R}^{T\times r_2}$ as the *factor matrices*, and $\boldsymbol{M} = (\cdots \boldsymbol{m}_{j_3} \cdots) \in \{0,1\}^{L\times r_3}$ as the *membership matrix* such that $(\boldsymbol{M})_{ij} = 1$ if the i-th treatment belongs to the j-th cluster.

Under this representation, the multi-linear ranks r_1, r_2, and r_3 are constrained to be no greater than the dimensions N, T, and L, respectively. In this model, $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ capture latent customer characteristics (like age, gender) and temporal patterns, respectively. The *membership matrix* corresponds to a *cluster label vector* $\boldsymbol{z} = (z_1, \ldots, z_L)$, with $z_l = j$ if and only if $(\boldsymbol{M})_{ij} = 1$. Thus, we use $\boldsymbol{M}$ and $\boldsymbol{z}$ interchangeably to denote the clustering structure for the third mode. The membership matrix $\boldsymbol{M}$ clusters treatments with similar effects, reducing the number of treatments. Lastly, the core tensor $\mathcal{S}$ indicates the interactions among these latent factors within each treatment cluster.

3.5. Weighted likelihood estimation for Θ

For parameter estimation, we propose a weighted maximum likelihood estimation. In specific, the log-likelihood function of Θ given the hazard probability model (2) is

$$\begin{aligned}l(\Theta) = \sum_{i,t,l} &\mathbf{1}(\mathcal{Y}_{i,t-1,l} = 1)\left[\mathcal{Y}_{i,t,l}\log\{f(\theta_{i,t,l})\}\right. \\ &\left.+(1-\mathcal{Y}_{i,t,l})\delta_i \log\{1 - f(\theta_{i,t,l})\}\right].\end{aligned}$$

However, the log-likelihood function is infeasible to compute directly due to the counterfactuals; specifically, only $N \times T$ realizations $\mathcal{Y}^{\text{obs}}$ is observable, leaving other parts of $\mathcal{Y}$ missing. This missing data is not completely random, as it is determined by the treatment mechanism, possibly confounded by indication, leading to potential biases in estimating Θ when relying only on the observed data. To mitigate this issue, we use the inverse probability treatment weighting (IPTW) based on the propensity scores $\pi(\boldsymbol{a} \mid \boldsymbol{X}_i)$

to adjust for confounding biases:

$$l(\Theta) = \sum_{i,t,l=(\boldsymbol{A}_i)_{10}} w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)\left[\mathcal{Y}_{i,t,l}\log\{f(\theta_{i,t,l})\}\right.$$
$$\left.+(1-\mathcal{Y}_{i,t,l})\delta_i\log\{1-f(\theta_{i,t,l})\}\right], \tag{4}$$

where $w_i = \pi\{(l)_2 \mid \boldsymbol{X}_i\}^{-1}$. This weighting effectively generates a pseudo-population where the missingness in the data is uniform.

Besides, we can exploit additional structural relationships between outcomes $\boldsymbol{Y}_i$ and covariates $\boldsymbol{X}_i$ to improve the estimation algorithm. For example, in the loading matrix $\boldsymbol{U}_1$, rows corresponding to customers with similar covariates might exhibit similarity. To utilize this insight, we proposed to decompose $\boldsymbol{U}_1$ into $\boldsymbol{U}_1 = \boldsymbol{X}\boldsymbol{u}_{10} + \boldsymbol{u}_{11}$ where $\boldsymbol{u}_{10}$ and $\boldsymbol{u}_{11}$ are matrices of unknown parameters. This covariate-assisted formulation of $\boldsymbol{U}_1$ can be easily incorporated by our projected gradient descent in Algorithm 1, which not only refines our understanding of the data structure but also potentially improves the precision of our estimation.

4. Algorithm

4.1. Propensity score estimation

Typically, the propensity scores $\pi(\boldsymbol{a} \mid \boldsymbol{X}_i)$ are unknown in the observational studies and thus require estimation. However, for a high-dimensional treatment vector ($L = 2^k$), likelihood-based estimation can yield nearly zero values and weighting by their inverse $\{\pi(\boldsymbol{a} \mid \boldsymbol{X}_i;\widehat{\alpha})\}^{-1}$ is unstable (Yang et al., 2016). To improve the stability, we adopt the covariate balance propensity score (CBPS) methodology in Imai & Ratkovic (2014). In particular, the following moment conditions are formulated for $j = 1, \cdots, k$ and, $a = 0, 1$:

$$\mathbb{E}\left[\rho_{i,j}\mathbf{1}(A_{i,j}=a)\boldsymbol{b}(\boldsymbol{X}_i)\right] = \mathbb{E}\left[\boldsymbol{b}(\boldsymbol{X}_i)\right], \tag{5}$$

where $\rho_{i,j}^{-1} = \mathbb{P}(A_{i,j} \mid \boldsymbol{X}_i)$ and $\boldsymbol{b}(\boldsymbol{X})$ is a set of arbitrary basis functions, which can be the first, second, and higher-order moments of $\boldsymbol{X}$. The key insight of (5) stems from the central role of the propensity scores, which balance the covariate distribution within the treatment groups in terms of the basis functions. To increase the stability of propensity score weighting, we propose to estimate the weights by minimizing the entropy balance function: $-\min_{\rho_{i,j}}\sum_{i=1}^N \rho_{i,j}\log\rho_{i,j}$,subject to $\rho_{i,j} \geq 0$, $\sum_{i=1}^N \rho_{i,j}\mathbf{1}(A_{i,j}=a) = 1$, and $\sum_{i=1}^N \rho_{i,j}\mathbf{1}(A_{i,j}=a)b(\boldsymbol{X}_i) = N^{-1}\sum_{i=1}^N b(\boldsymbol{X}_i)$. The loss function is the entropy function of the weights that enforces the weights to be as close to one as possible, which reduces the variability due to heterogeneous weights (Hainmueller, 2012; Lee et al., 2022; 2023). The final propensity score weights for the treatment vector become $\widehat{w}_i = \prod_{j=1}^k \widehat{\rho}_{i,j}$. Challenges may arise when we are facing a large number of moment constraints, which increases the chance of conflicting restrictions and thus do not produce a feasible solution space. One remedy is to couple the objective function (i.e., entropy balance) with regularization on the moment constraints to carefully select the important subset for balancing (Ning et al., 2020).

4.2. Projected gradient descent and spectral clustering

To maximize the weighted log-likelihood function (4) with w_i replaced by $\widehat{w}_i$ under model (3), we use the projected gradient descent method to obtain $(\widehat{\mathcal{S}}, \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2)$ along with the spectral clustering on the third mode to find the optimal membership $\widehat{\boldsymbol{M}}$. In particular, we first compute the partial gradient of $l(\Theta)$ with respect to $(\mathcal{S}, \boldsymbol{U}_1, \boldsymbol{U}_2)$:

$$\frac{\partial l(\Theta)}{\partial \mathcal{S}} = \nabla l \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal \times_3 \boldsymbol{M}^\intercal,$$
$$\frac{\partial l(\Theta)}{\partial \boldsymbol{U}_1} = \mathcal{M}_{(1)}(\nabla l)(\boldsymbol{U}_2 \otimes \boldsymbol{M})\mathcal{M}_{(1)}(\mathcal{S})^\intercal,$$
$$\frac{\partial l(\Theta)}{\partial \boldsymbol{U}_2} = \mathcal{M}_{(2)}(\nabla l)(\boldsymbol{U}_1 \otimes \boldsymbol{M})\mathcal{M}_{(2)}(\mathcal{S})^\intercal,$$

where $\nabla l = \partial l(\Theta)/\partial\Theta$. Next, we update the current solution $(\widehat{\mathcal{S}}, \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2)$ by subtracting $\eta \cdot \partial l(\Theta)/\partial(\widehat{\mathcal{S}}, \widehat{\boldsymbol{U}}_1, \widehat{\boldsymbol{U}}_2)$, which moves it towards the opposite direction of partial gradients with step size η. To find the optimal membership for the third mode, we perform the nearest-neighbor search to update our estimate for the clustering labels z:

$$\widehat{z}_l = \arg\min_{b\in[r_3]} \|\mathcal{M}_{(3)}(\widehat{\mathcal{F}})_{l,:} - \mathcal{M}_{(3)}(\widehat{\mathcal{S}})_{b,:}\|_2^2,$$

where $l = 1, \cdots, L$, $\widehat{\mathcal{F}} = \widehat{\Theta} \times_1 (\widehat{\boldsymbol{U}}_1)^\intercal \times_2 (\widehat{\boldsymbol{U}}_2)^\intercal$ is the projected mode-3 slices, $\widehat{\mathcal{S}} = \widehat{\Theta} \times_1 (\widehat{\boldsymbol{U}}_1)^\intercal \times_2 (\widehat{\boldsymbol{U}}_2)^\intercal \times_3 (\widehat{\boldsymbol{W}})^\intercal$ is the projected mode-3 block means, and $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{M}}(\mathrm{diag}(\mathbf{1}_L^\intercal\widehat{\boldsymbol{M}}))^{-1}$. Intuitively, $\mathcal{M}_{(3)}(\widehat{\mathcal{F}})_{l,:}$ contains the information of the l-th treatment, and $\mathcal{M}_{(3)}(\widehat{\mathcal{S}})_{b,:}$ contains the information of the b-th cluster. Our strategy is to find b for each l such that $\mathcal{M}_{(3)}(\widehat{\mathcal{S}})_{b,:}$ is the closest to $\mathcal{M}_{(3)}(\widehat{\mathcal{F}})_{l,:}$. Besides, $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{S}}$ utilize the information from the other two modes for projection, which can significantly reduce the noise level within the estimated parameter tensor $\widehat{\Theta}$.

We provide the details of our procedure for optimizing (4) in Algorithm 1. In practice, we recommend a BIC-type criterion to select the rank parameters (r_1, r_2, r_3), and all the tuning parameters are tuned sequentially (Ibriga & Sun, 2023).

5. Statistical theory

In this section, we study the statistical properties of $\widehat{\Theta}$ under the tensorized latent factor block hazard model (3). Mainly,

Algorithm 1 Projected gradient descent and spectral clustering for minimizing (4)

Input: Observed data tuple $\mathbf{V}_i = (\boldsymbol{X}_i, \boldsymbol{A}_i, \delta_i, \boldsymbol{Y}_i)$, estimated weights $\widehat{w}$, stepsize η, ranks r_1, r_2 and r_3
//Initialization
Initialize the parameter tensor estimator $\widehat{\Theta}^{(0)}$ via the logistic regression classifier
for $k = 1, 2$ **do**
  Initialize the singular space estimator for the first two modes via $\widehat{\boldsymbol{U}}_k^{(0)} = \mathrm{SVD}_{r_k}\{\mathcal{M}_{(k)}(\widehat{\Theta}^{(0)})\}$
end for
Compute $\widehat{\boldsymbol{F}}^{(0)} = \mathcal{M}_{(3)}(\widehat{\Theta}^{(0)})(\widehat{\boldsymbol{U}}_1^{(0)} \otimes \widehat{\boldsymbol{U}}_2^{(0)})$
Find $\widehat{\boldsymbol{z}}^{(0)} \in [r_3]^L$ and centroids $\widehat{x}_1, \cdots, \widehat{x}_{r_3} \in \mathbb{R}^{NT}$ such that $\sum_{l=1}^L \|(\widehat{\boldsymbol{F}})_{l:}^\intercal - \widehat{x}_{\widehat{\boldsymbol{z}}_l^{(0)}}\|_2^2 \leq \min_{x,\boldsymbol{z}} \sum_{l=1}^L \|(\widehat{\boldsymbol{F}})_{l:}^\intercal - x_{\boldsymbol{z}_l}\|_2^2$
Compute $\widehat{\boldsymbol{W}}^{(0)} = \widehat{\boldsymbol{M}}^{(0)}(\mathrm{diag}(\boldsymbol{1}_L^\intercal \widehat{\boldsymbol{M}}^{(0)}))^{-1}$, where $\widehat{\boldsymbol{M}}^{(0)}$ is determined by $\widehat{\boldsymbol{z}}^{(0)}$
Initialize the core tensor $\widehat{\mathcal{S}}^{(0)} = \widehat{\Theta}^{(0)} \times_1 (\widehat{\boldsymbol{U}}_1^{(0)})^\intercal \times_2 (\widehat{\boldsymbol{U}}_2^{(0)})^\intercal \times_3 (\widehat{\boldsymbol{W}}^{(0)})^\intercal$
//Updating
for $I = 1, \cdots, I_{\max}$ **do**
  for $k = 1, 2$ **do**
    $\widehat{\boldsymbol{U}}_k^{(I)} = \widehat{\boldsymbol{U}}_k^{(I-1)} - \eta \frac{\partial L(\mathcal{S}^{(I-1)}, \widehat{\boldsymbol{U}}_1^{(I-1)}, \widehat{\boldsymbol{U}}_2^{(I-1)}, \widehat{\boldsymbol{M}}^{(I-1)})}{\partial \boldsymbol{U}_k}$
    $\widehat{\boldsymbol{U}}_k^{(I)} = \boldsymbol{P}_k(\widehat{\boldsymbol{U}}_k^{(I)})$, where $\boldsymbol{P}_k(\cdot)$ is the projection operator for $\boldsymbol{U}_k$
  end for
  Update $\widehat{\boldsymbol{F}}^{(I)}$, $\widehat{\boldsymbol{z}}^{(I)}$, $\widehat{\boldsymbol{M}}^{(I)}$, and $\widehat{\boldsymbol{W}}^{(I)}$ via the nearest-neighbor search
  Update $\widehat{\mathcal{S}}^{(I)} = \widehat{\mathcal{S}}^{(I-1)} - \eta \frac{\partial L(\mathcal{S}^{(I-1)}, \widehat{\boldsymbol{U}}_1^{(I)}, \widehat{\boldsymbol{U}}_2^{(I)}, \widehat{\boldsymbol{M}}^{(I)})}{\partial \widehat{\mathcal{S}}^{(I-1)}}$
end for
//Final results
$\widehat{\mathcal{S}} = \widehat{\mathcal{S}}^{(I_{\max})}$, $\widehat{\boldsymbol{U}}_1 = \widehat{\boldsymbol{U}}_1^{(I_{\max})}$, $\widehat{\boldsymbol{U}}_2 = \widehat{\boldsymbol{U}}_2^{(I_{\max})}$ and $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}^{(I_{\max})}$
Output: Estimated the parameter tensor $\widehat{\Theta} = \widehat{\mathcal{S}} \times_1 \widehat{\boldsymbol{U}}_1 \times_2 \widehat{\boldsymbol{U}}_2 \times_3 \widehat{\boldsymbol{M}}$

we focus on evaluating the performance in two metrics: 1) estimation: the estimation accuracy for the parameter tensor Θ (Section 5.1); 2) clustering: the correct recovery rate of the membership $\boldsymbol{M}$ (Section 5.2). To begin with, we provide a summary of the notation in Table 1.

Notation	Definition
$\mathcal{F}$	$\Theta \times_1 (\boldsymbol{U}_1)^\intercal \times_2 (\boldsymbol{U}_2)^\intercal$
$\mathcal{S}$	$\widehat{\Theta} \times_1 (\boldsymbol{U}_1)^\intercal \times_2 (\boldsymbol{U}_2)^\intercal \times_3 (\boldsymbol{W})^\intercal$
$\boldsymbol{W}$	$\boldsymbol{M}(\mathrm{diag}(\boldsymbol{1}_L^\intercal \boldsymbol{M}))^{-1}$
$p_{\min}$	$p_{\min} \leq \mathbb{P}(\delta_i = 1, \mathcal{Y}_{i,t,l} = 1)\ \forall\, i, t$ and l
$w_{\min}, w_{\max}$	$w_{\min} \leq \pi(\boldsymbol{a} \mid \boldsymbol{X}_i)^{-1} \leq w_{\max}$
r_1, r_2	number of latent unit and temporal factors
r_3	number of groups for the treatments

Table 1. Summary of notation

5.1. Upper bound error for estimation

The estimation accuracy of $\widehat{\Theta}$ is evaluated by its deviation to Θ in Frobenius norm. First, we define two quantities L_α and γ_α to control the steepness and convexity of the link function $f(\cdot)$, where $L_\alpha = \sup_{|\theta| \leq \alpha} [f'(\theta)/f(\theta), f'(\theta)/\{1 - f(\theta)\}]$ and $\gamma_\alpha = \inf_{|\theta| \leq \alpha} [\{f'(\theta)\}^2/f^2(\theta) - f''(\theta)/f(\theta), f''(\theta)/\{1 - f(\theta)\} + \{f'(\theta)\}^2/\{1 - f(\theta)\}^2]$, where $f'(\theta) = df(\theta)/d\theta$.

Theorem 5.1 establishes the upper bound of the estimation error of $\widehat{\Theta}$.

Theorem 5.1. *Under Assumptions A1) to A4) and some regularity conditions, suppose $\mathcal{Y}$ is the binary tensor characterized by the parameter tensor Θ as model (2) with the link function $f(\cdot)$. Let $\widehat{\Theta}$ be the local maximizer of (4), there exist constants c_0, C_0, C_1 and C_2, such that with probability greater than $1 - c_0(N + T + L)^{-2}$, we have*

$$\frac{\|\widehat{\Theta} - \Theta\|_F^2}{NT} \leq C_0 \frac{\|\Theta\|_{\max}^2}{N p_{\min}} \log(N + T + L)$$
$$\vee\, C_1 \frac{L_\alpha^2}{\gamma_\alpha^2} \frac{w_{\max}^2}{w_{\min}^2} \frac{r_1 r_2 r_3 (N \vee T)}{N T p_{\min}^2 \max(r_1, r_2, r_3)} \log^2(N + T + L)$$
$$\vee\, C_2 \frac{r_1 r_2 r_3 (N \vee T) \|\Theta\|_{\max}^2}{N T p_{\min}^2 \max(r_1, r_2, r_3)} \log(N + T + L).$$

Theorem 5.1 shows that the upper bound for estimation error converges to zero as the sample size N increases; see Appendix for the experimental evidence. The parameter $p_{\min}$ plays an important role in our theoretical analysis as the recovery of Θ will be harder when $p_{\min}$ becomes smaller. Intuitively, if $p_{\min}$ is too small, it is unlikely to observe the churn statuses for all the customers and thus recovering their churn patterns will be impossible. Furthermore, the estimation error for the survival probabilities is also bounded if $f(\cdot)$ satisfies the local Lipschitz condition, that

is, $\|f(\widehat{\Theta}) - f(\Theta)\|_F \le c_1 \|\widehat{\Theta} - \Theta\|_F$ for some constant c_1. Below, we present a special case of Theorem 5.1 under the logistic link function.

Corollary 5.2. *Assume the logistic model for $\mathcal{Y}_{i,t,l}$ with link function $f(\theta) = e^{\theta/\sigma}/(1+e^{\theta/\sigma})$, it can be shown that*

$$L_\alpha^2/\gamma_\alpha^2 = \frac{(1+e^{\alpha/\sigma})^4\sigma^2}{e^{2\alpha/\sigma}} = \left(2 + e^{\alpha/\sigma} + \frac{1}{e^{\alpha/\sigma}}\right)^2 \sigma^2.$$

Further, suppose $w_{\min} \asymp w_{\max}$, $p_{\min}$ is bounded and $\|\Theta\|_{\max} \asymp \sigma$, the estimation error in Theorem 5.1 becomes

$$\frac{\|\widehat{\Theta} - \Theta\|_F^2}{NT} \lesssim \frac{r_1 r_2 r_3 (N \vee T)}{NT \max(r_1, r_2, r_3)} \times (\sigma^2 \vee \|\Theta\|_{\max}^2) poly \log(N+T+L), \tag{6}$$

where poly $\log(\cdot)$ is a certain polynomial of the logarithmic function. This non-asymptotic bound (6) is similar to the risk bound in Theorem 3, Xia et al. (2021).

5.2. Upper bound error for clustering

Next, we examine the clustering accuracy of our proposed model for the third mode. The most common metric to assess the clustering performance is the *classification error rate*, defined by $h(\boldsymbol{c}, \boldsymbol{d}) = \min_{\pi \in \Pi_{r_3}} \sum_{l=1}^L \mathbf{1}\{\boldsymbol{c}_l \neq \pi(\boldsymbol{d})_l\}/L$, for two label vectors $\boldsymbol{c} = (c_1, \cdots, c_L)^\intercal$ and $\boldsymbol{d} = (d_1, \cdots, d_L)^\intercal$, where Π_{r_3} is the collection of all permutations of $\{1, \cdots, r_3\}$. Let $\widehat{\boldsymbol{z}}$ be the estimated clustering labels, we claim that $\widehat{\boldsymbol{z}}$ is a consistent clustering for $\boldsymbol{z}$ if $\mathbb{P}\{h(\widehat{\boldsymbol{z}}, \boldsymbol{z}) > \varepsilon\} \to 0$ when the sample size N goes to infinity for any $\varepsilon > 0$. In order to facilitate our theoretical analysis, we introduce the concept of *misclassification loss*:

$$g(\boldsymbol{c}, \boldsymbol{d}) = \min_{\pi \in \Pi_{r_3}} \frac{1}{L} \sum_{l=1}^L \|\mathcal{M}_{(3)}(\mathcal{S})_{(\boldsymbol{c})_l,:} - \mathcal{M}_{(3)}(\mathcal{S})_{\pi(\boldsymbol{d})_l,:}\|_2^2,$$

which is a more convenient measure compared to $h(\boldsymbol{c}, \boldsymbol{d})$. Moreover, a close relationship between $h(\boldsymbol{c}, \boldsymbol{d})$ and $g(\boldsymbol{c}, \boldsymbol{d})$ can be formulated as $h(\boldsymbol{c}, \boldsymbol{d}) \le g(\boldsymbol{c}, \boldsymbol{d})/\Delta_{\min}^2$ (Lemma 1, Han et al. (2022)), which implies that it suffices to bound $g(\boldsymbol{c}, \boldsymbol{d})$ for establishing the clustering consistency.

Theorem 5.3. *Under the same condition in Theorem 5.1. Assume the signal-to-noise ratio (SNR) satisfies*

$$SNR = \frac{\Delta_{\min}^2}{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)} \gtrsim \frac{w_{\max}^2 (N \vee T) r_1^2 r_2^2 r_3 \mu_0^2 \log^2(N+T+L)}{w_{\min}^2 p_{\min}^2 NTL \max(r_1, r_2, r_3)},$$

where $\Delta_{\min}$ measures the minimum separation of the projected block means, and μ_0 measures the incoherence of Θ; see details in the Appendix. There exists a constant c_0, such that, with probability greater than $1 - c_0(N+T+L)^{-2}$,

$$h(\widehat{z}, z) \le g(\widehat{z}, z)/\Delta_{\min}^2 \lesssim \frac{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)}{\Delta_{\min}^2 (N+T+L)^2}.$$

The notion of SNR under the proposed model is quantified by the minimum gaps between the projected means on the third mode, i.e., $\Delta_{\min}$, over the level of noises induced by the parameter tensor and the Bernoulli model (2). If the SNR is large enough as stated in Theorem 5.3, the consistency of the clustering will be established; see Appendix for the experimental evidence. A similar SNR condition also appears in Theorem 2, Han et al. (2022).

6. Numerical studies

In this section, we examine the performance of our proposed model under various settings. For starter, the baseline covariates $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ are generated by $\boldsymbol{X}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ with $d = 3$. We generate the true parameter tensor Θ with each entry $\theta_{i,t,l}$ defined by: $\theta_{i,t,l} = (\boldsymbol{X}_i^\intercal \eta_N) \cdot (t\eta_T/T) \cdot \text{cum}\{(l)_2\}\eta_L$, where $\eta_N = (1,1,1)$, $\eta_T = 1$, $\eta_L = 1$, and $\text{cum}\{(l)_2\}$ indicates the number of active treatments in $(l)_2$. The potential outcome $\mathcal{Y}_{i,t,l}$ is generated sequentially with the conditional probability $P(\mathcal{Y}_{i,t,l} = 1 \mid \mathcal{Y}_{i,t-1,l} = 1, \boldsymbol{X}_i) = \text{expit}(\theta_{i,t,l})$ when $\mathcal{Y}_{i,t-1,l} = 1$ and $\boldsymbol{A}_i = (l)_2$; otherwise, $\mathcal{Y}_{i,t,l} = 0$ by definition. The lifetime for each customer i is computed by $\text{Time}_i = \sum_{t=1}^T \sum_{l=1}^L \mathbf{1}\{\boldsymbol{A}_i = (l)_2\}\mathcal{Y}_{i,t,l}$, and we randomly select 20% patients to be right-censored with random censoring time uniformly drawn from $[0, \text{Time}_i]$. Next, each entry of the k-dimensional binary treatment vector $\boldsymbol{A}_i$ is generated independently for $j = 1, \cdots, k$:

$$A_{i,j} \mid \boldsymbol{X}_i \sim \text{Bernoulli}\left\{\frac{\exp(\alpha_A^\intercal \boldsymbol{X}_i)}{1 + \exp(\alpha_A^\intercal \boldsymbol{X}_i)}\right\},$$

where $\alpha_A = (.5, .5, .5)$. The CBPS method in Section 4.1 is utilized for propensity score estimation with the basis functions $b(\boldsymbol{X})$ being the first moment of $\boldsymbol{X}$. We consider the setting where $T = 5, 10$, $N = 100, 300, 500, 1000, 2000$, and $k = 2, 3, 4$. All simulation results are reported based on 100 data replications.

One primary goal for the customer churn analysis is to identify the optimal treatment that leads to the longest customer retention time. With the estimated parameter tensor $\widehat{\Theta}$, one can find the individual optimal treatment $\hat{D}_{i,\text{opt}}$ by $\max_l \sum_{t=1}^T \prod_{s=1}^t \text{expit}(\widehat{\theta}_{i,s,l})$, which maximize the estimated expected lifetime for each customer i. We showcase the effectiveness of our proposed model for identifying the optimal treatment compared with other methods. In particular, we consider the competitive methods within two categories. 1) binary classification: logistic regression classifier (logit), random forest classifier (RF), neural network classifier (NN), support vector machine (SVM), gradient-boosted classifier (gradBoost), adaptive boosting classifier (AdaBoost), and a soft voting classifier combining all mentioned binary classifiers (Vote); 2) survival models: Cox proportional hazard model (Cox-PH), survival random forest

(surv-RF), gradient-boosted Cox PH model with regression trees as base learners (Cox-modelBoost), gradient boosting with component-wise least squares as base learners (Cox-gradBoost). The hyperparameters for these algorithms are chosen by default from packages `sklearn` and `sksurv`.

We assess the performance by the cumulative regret and decision accuracy using the true optimal treatment $D_{i,\text{opt}}$; more definitions are deferred to the Appendix. Specifically, cumulative regret is defined as the average difference between the survival probabilities evaluated under the estimated and the true optimal treatments summing over all the time points. Therefore, the regret captures how the estimated incorrect decisions impair the outcome. The decision accuracy describes the proportion of estimated optimal treatment that matches the true optimal treatment. We evaluate different methods by their cumulative regrets and decision accuracy (Figure 2) for varying N, T and k. One can observe the proposed model outperforms the other methods by large margins in all cases.

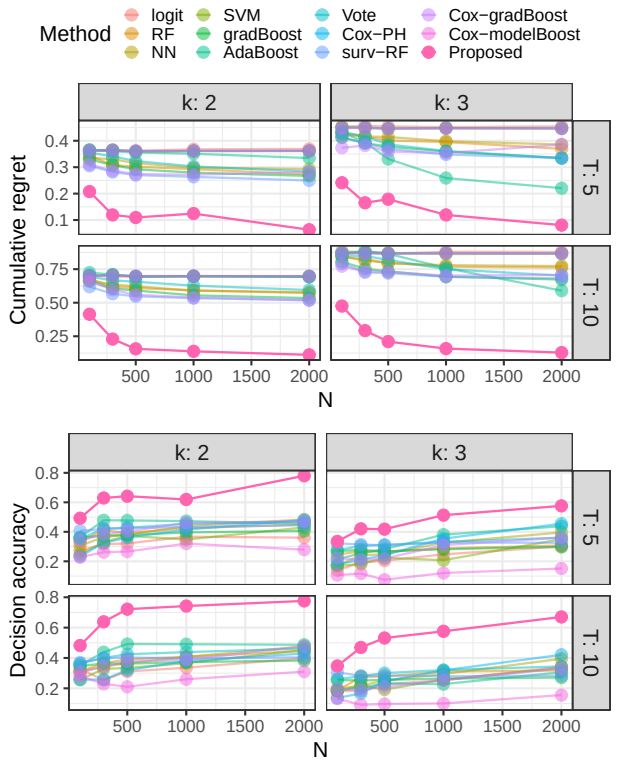

Figure 2. Cumulative regret (top) and decision accuracy (bottom) of the proposed method and other competitors when $N = 100, 300, 500, 1000, 2000$, $T = 5, 10$ and $k = 2, 3$.

7. Real-data application

In this section, we apply our proposed method to one bank customer churn data from a Kaggle competition[1]. Four important indicators of customer loyalty are considered: card types (0 for Silver and Gold, and 1 for Platinum and Diamond), the number of bank products (0 for only one product, and 1 for more than one), whether the customer has complained or not, and the post-complaint scores (0 for rates less than 3, and 1 otherwise). It is known that companies might employ different interventions based on different subgroups of customers to avoid customer churn. Therefore, we formulate the binary treatment vector $\boldsymbol{A}_i$ according to these indicators. Next, we consider eight customer characteristics $\boldsymbol{X}_i$, including age, gender, geography, estimated salary, account balance, earned credit points, and two customer characteristic indicators: whether or not the customers have credit cards and whether or not the customers are active. The customer retention time Y_i is measured by the number of years that the customers have engaged with the bank and the churn status δ_i is an indicator of whether or not the customers left the bank.

For the model training, we divide the dataset into 80% training data and 20% test data, implementing 5-fold cross-validation. Continuous variables are standardized, and categorical variables are one-hot encoded. The goodness-of-fit metrics for assessing different methods are the concordance index (C-index) and the average time-dependent area under the curve (AUC). C-index, a widely used index for survival analysis, examines the performance by the fraction of pairs whose predicted retention times have the correct order compared to their observed retention times in the test set. The time-dependent ROC curve evaluates the model's ability to distinguish the customers who exit by a given time t from those who exit after t. Table 2 compares the results of our method with other methods in terms of their goodness-of-fit. In conclusion, our proposal performs well, exhibiting notably superior performance compared to other methods.

To better explore the homogeneity of the intervention effect, we estimate the expected customer lifetime within each intervention group under our model. Figure 3 shows that the customers with more bank products tend to have longer retention times. This finding aligns with our expectations, as customers associated with more bank products are more loyal and inclined to remain engaged with the company. Conversely, customers who have fewer bank products and complained previously exhibit the shortest customer lifetimes regardless of their post-complaint scores. Hence, these customers are more likely to leave the company. Given the higher cost involved in acquiring new customers rather than retaining the existing ones, our churn analysis alerts

[1]https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn.

	C-index (↑)	average AUC (↑)
logit	0.19 ± 0.08	0.43 ± 0.04
RF	0.44 ± 0.01	0.46 ± 0.01
NN	0.31 ± 0.02	0.36 ± 0.01
SVM	0.49 ± 0.04	0.47 ± 0.12
gradBoost	0.50 ± 0.01	0.47 ± 0.02
AdaBoost	0.45 ± 0.03	0.39 ± 0.04
Vote	0.44 ± 0.01	0.39 ± 0.03
Cox-PH	0.30 ± 0.01	0.30 ± 0.01
surv-RF	0.46 ± 0.01	0.40 ± 0.01
Cox-gradBoost	0.43 ± 0.01	0.40 ± 0.01
Cox-modelBoost	0.46 ± 0.01	0.42 ± 0.01
Proposed	$\mathbf{0.58} \pm 0.02$	$\mathbf{0.57} \pm 0.02$

Table 2. Evaluation metrics on the bank customer churn data.

the need for the company to design retention campaigns targeting the customers with fewer bank products who have raised complaints before.

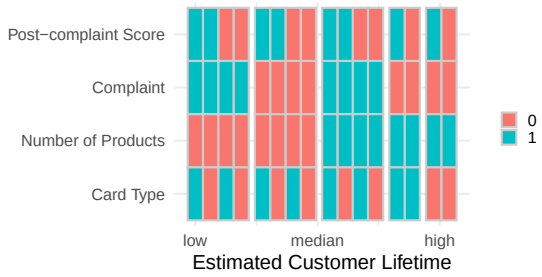

Figure 3. Estimated intervention structure by the proposed model for the bank customer churn data. All interventions are clustered by blocks and ordered by their expected customer lifetimes.

8. Discussion

In this paper, we focus on the causal analysis of the customer churn problem with multiple interventions. In particular, we adopt the idea of 1-bit tensor completion to estimate the survival probabilities under all interventions. Moreover, we propose the tensorized latent factor block hazard model to cluster the interventions with similar impacts. This model enables us to identify the optimal intervention group, which improves the practicality of implementing the optimal retention strategies in practice. The proposed method can be extended into several aspects. First, we only consider the time-invariant treatment in this paper. When treatment is time-varying, two classes of models namely marginal structural models (Yang et al., 2018) and structural failure time models (Yang et al., 2020) are useful, which, however, often posit parametric structural model assumptions. The proposed framework can be extended to this setting by expanding the treatment mode. Second, our current identification assumption requires the treatment ignorability in the sense that all confounders are captured and adjusted. One interesting future direction is to extend the current framework under the latent ignorability of treatment assignment in the sense that treatment ignorability holds when conditioning on the latent factors (Lewis & Syrgkanis, 2021; Agarwal & Syrgkanis, 2022).

Software and Data

Our Python codes with illustrative examples are available at `https://github.com/Gaochenyin/Low-Rank-Tensor-Block-Hazard-Model`

Acknowledgements

We would like to thank the anonymous (meta-)reviewers of ICML 2024 for helpful comments. This work is partially supported by the U.S. National Science Foundation and National Institute of Health.

Impact Statement

This paper presents work aimed at predicting customer churn and developing informed strategies to improve customer retention. The potential societal consequences of this work could be significant, including fostering more sustainable business practices, enhancing customer satisfaction, promoting economic stability by reducing the frequency of business failures, and contributing to higher levels of service quality and consumer trust in various industries.

References

Agarwal, A. and Syrgkanis, V. Synthetic blip effects: Generalizing synthetic controls for the dynamic treatment regime. *arXiv preprint arXiv:2210.11003*, 2022.

Agarwal, A., Shah, D., and Shen, D. Synthetic interventions. *arXiv preprint arXiv:2006.07691*, 2020.

Agarwal, A., Dahleh, M., Shah, D., and Shen, D. Causal matrix completion. *arXiv preprint arXiv:2109.15154*, 2021.

Ashraphijuo, M. and Wang, X. Union of low-rank tensor spaces: Clustering and completion. *Journal of Machine Learning Research*, 21:1–36, 2020.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116:1–15, 2021.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25:890–896, 2009.

Buckinx, W., Verstraeten, G., and Van den Poel, D. Predicting customer loyalty using the internal transactional database. *Expert systems with applications*, 32:125–134, 2007.

Cai, C., Li, G., Poor, H. V., and Chen, Y. Nonconvex low-rank tensor completion from noisy data. *Operations Research*, 70:1–19, 2021.

Cai, T. and Zhou, W.-X. A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14:3619–3647, 2013.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Cao, Y., Zhang, A., and Li, H. Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, 107:75–92, 2020.

Ching, T., Zhu, X., and Garmire, L. X. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14:e1006076, 2018.

Coussement, K. and Van den Poel, D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, 34:313–327, 2008.

Davenport, M. A., Plan, Y., Van Den Berg, E., and Wootters, M. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3:189–223, 2014.

Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27:025010, 2011.

Gao, C. and Zhang, A. Y. Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50:1066–1094, 2022.

Ghadermarzy, N., Plan, Y., and Yilmaz, O. Learning tensors from partial binary measurements. *IEEE Transactions on Signal Processing*, 67:29–40, 2018.

Hainmueller, J. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20: 25–46, 2012.

Han, R., Luo, Y., Wang, M., and Zhang, A. R. Exact clustering in tensor block model: Statistical optimality and computational limit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1666–1698, 2022.

Ibriga, H. S. and Sun, W. W. Covariate-assisted sparse tensor completion. *Journal of the American Statistical Association*, 118:2605–2619, 2023.

Imai, K. and Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76:243–263, 2014.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. Random survival forests. *The Annals of Applied Statistics*, 2:841 – 860, 2008. doi: 10.1214/08-AOAS169. URL https://doi.org/10.1214/08-AOAS169.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18: 1–12, 2018.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51:455–500, 2009.

Laber, E. B., Lizotte, D. J., and Ferguson, B. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70:53–61, 2014.

Larivière, B. and Van den Poel, D. Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, 27: 277–285, 2004.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Lee, D., Yang, S., and Wang, X. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10:415–440, 2022.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. Improving trial generalizability using observational studies. *Biometrics*, 79:1213–1225, 2023.

Lewis, G. and Syrgkanis, V. Double/debiased machine learning for dynamic treatment effects. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22695–22707, 2021.

Liu, J., Musialski, P., Wonka, P., and Ye, J. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35:208–220, 2012.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, 37:3776–3788, 2018.

Lu, J. Predicting customer churn in the telecommunications industry—-an application of survival analysis modeling using sas. *SAS User Group International (SUGI27) Online Proceedings*, 114:27, 2002.

Lu, N., Lin, H., Lu, J., and Zhang, G. A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, 10:1659–1665, 2012.

Luo, Z., Qi, L., and Toint, P. L. Tensor bernstein concentration inequalities with an application to sample estimators for high-order moments. *Frontiers of Mathematics in China*, 15:367–384, 2020.

Ma, H., Zeng, D., and Liu, Y. Learning individualized treatment rules with many treatments: A supervised clustering approach using adaptive fusion. *Advances in Neural Information Processing Systems*, 35:15956–15969, 2022.

Ma, Z., Ma, Z., and Yuan, H. Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21:86–152, 2020.

Mandal, D. and Parkes, D. Weighted tensor completion for time-series causal inference. *arXiv preprint arXiv:1902.04646*, 2019.

Mao, X., Wang, Z., and Yang, S. Matrix completion under complex survey sampling. *Annals of the Institute of Statistical Mathematics*, 75:463–492, 2023.

Mao, X., Wang, H., Wang, Z., and Yang, S. Mixed matrix completion in complex survey sampling under heterogeneous missingness. *Journal of Computational and Graphical Statistics*, pp. 1–19, 2024.

Massart, P. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28:863–884, 2000.

Maystre, L. and Russo, D. Temporally-consistent survival analysis. *Advances in Neural Information Processing Systems*, 35:10671–10683, 2022.

Mishra, A. and Reddy, U. S. A novel approach for churn prediction using deep learning. In *2017 IEEE international conference on computational intelligence and computing research (ICCIC)*, pp. 1–4. IEEE, 2017.

Ning, Y., Sida, P., and Imai, K. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107:533–554, 2020.

Pan, Y. and Zhao, Y.-Q. Improved doubly robust estimation in learning optimal individualized treatment rules. *Journal of the American Statistical Association*, 116:283–294, 2021.

Pearl, J. *Causal inference in statistics: An overview*. John Wiley & Sons, 2009.

Qian, M. and Murphy, S. A. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39: 1180, 2011.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688, 1974.

Rudd, D. H., Huo, H., and Xu, G. Causal analysis of customer churn using deep learning. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pp. 319–324. IEEE, 2021.

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

Tsiatis, A. A. *Semiparametric theory and missing data*. Springer, 2006.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

Umayaparvathi, V. and Iyakutti, K. Automated feature selection and churn prediction using deep learning models. *International Research Journal of Engineering and Technology (IRJET)*, 4:1846–1854, 2017.

Van der Laan, M. J. and Rose, S. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Wang, M. and Li, L. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research*, 21, 2020.

Xia, D. and Yuan, M. On polynomial time methods for exact low rank tensor completion. *arXiv preprint arXiv:1702.06980*, 2017.

Xia, D., Yuan, M., and Zhang, C.-H. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, 49:76–99, 2021.

Xie, Y., Li, X., Ngai, E., and Ying, W. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36:5445–5449, 2009.

Yang, S. Semiparametric Estimation of Structural Nested Mean Models with Irregularly Spaced Longitudinal Observations. *Biometrics*, 78:937–949, 04 2021.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72:1055–1065, 2016.

Yang, S., Tsiatis, A. A., and Blazing, M. Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach. *Biometrics*, 74:900–909, 2018.

Yang, S., Pieper, K., and Cools, F. Semiparametric estimation of structural failure time models in continuous-time processes. *Biometrika*, 107:123–136, 2020.

Yu, Y., Wang, T., and Samworth, R. J. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102:315–323, 2015.

Zhang, R., Li, W., Tan, W., and Mo, T. Deep and shallow model for insurance churn prediction service. In *2017 IEEE International Conference on Services Computing (SCC)*, pp. 346–353. IEEE, 2017.

Zhao, L. and Feng, D. Deep neural networks for survival analysis using pseudo values. *IEEE journal of biomedical and health informatics*, 24:3308–3314, 2020.

Zhu, X., Yao, J., and Huang, J. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 544–547. IEEE, 2016.

A. Additional Numerical Experiments

A.1. Convergence analysis for estimation and clustering

We first assess the effectiveness of the proposed model in recovering the parameter tensor Θ. The performance is evaluated by the normalized tensor mean squared error as $\ell_2(\widehat{\Theta}) = \|\widehat{\Theta} - \Theta\|_F^2 / \|\Theta\|_F^2$, and the classification error rate. Figure A1 shows that the estimation and clustering errors decrease as the sample size N increases. This shows that larger sample sizes enhance the performance of the proposed model, which corroborates our theoretical results in Theorems 5.1 and 5.3.

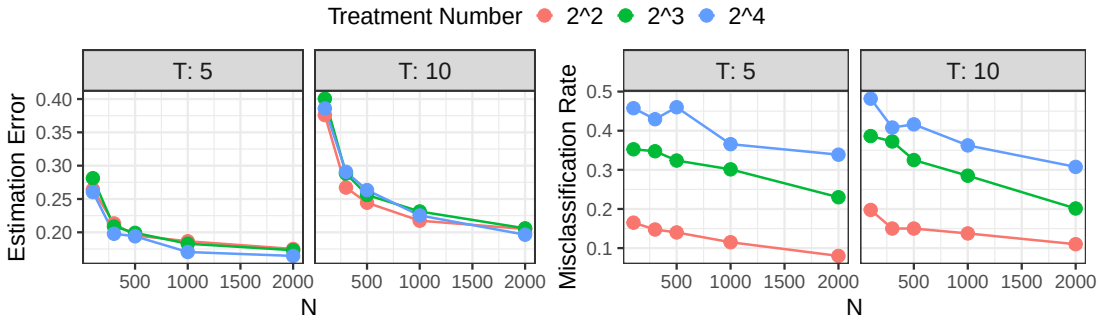

Figure A1. Normalized mean squared error (left) and the misclassification error rate (right) when $T = 5$ and $N = 100, 300, 500, 1000, 2000$ over 100 data replications.

Since the parameter tensor Θ is estimated, the average survival function $S^{(\boldsymbol{a})}(t)$ under any treatment can be obtained by $\widehat{S}^{(\boldsymbol{a})}(t) = N^{-1} \sum_{i=1}^{N} \prod_{s=1}^{t} \mathrm{expit}(\widehat{\theta}_{i,s,l})$. In particular, we plot the average estimated survival functions for all treatments over100 data replications when $N = 1000$ to assess the clustering results in Figure A2. The results imply that the more active treatments one receives, the higher the survival probabilities will be, which aligns with our data generation process of $\theta_{i,t,l}$ as it relies on the number of active treatments $\mathrm{cum}\{(l)_2\}$.

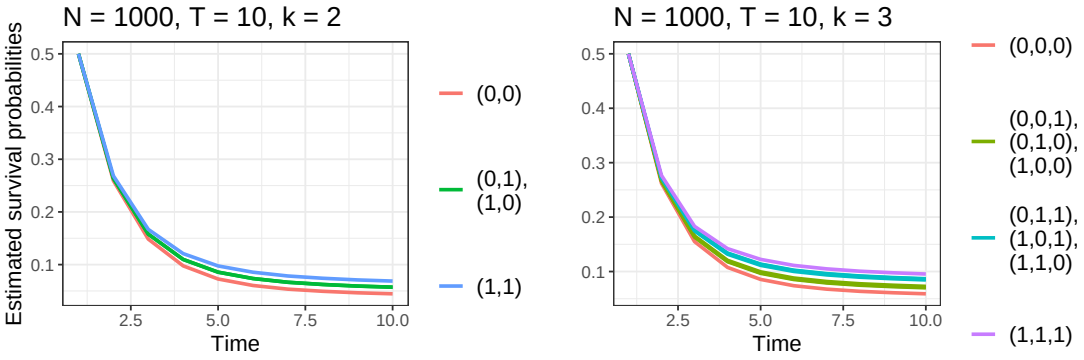

Figure A2. Plot of the average estimated survival probabilities for all treatments when $N = 1000, T = 10, k = 2$ (left) and $k = 3$ (right) over 100 data replications.

Next, we provide the complexity of the algorithm by presenting the running time and the normalized tensor mean squared errors ℓ_2 for our numerical experiments. Empirically, we present the results in Table A1 for both the 10000-iteration and 1000-iteration projected gradient descent algorithms. In summary, the running time of the proposed framework is linear with respect to N, T, k, and the number of iterations, which aligns with the theory for the gradient descent algorithm. Additionally, we suggest using the 1000-iteration projected gradient descent if saving time is a priority (the running time is significantly reduced), as the performance, in terms of normalized tensor mean squared error, does not decrease significantly when reducing the number of iterations from 10000 to 1000. All experiments are conducted on a computer with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz and 32GB RAM.

N	T	k	running time$^{(10000)}$	running time$^{(1000)}$	$\ell_2^{(10000)}$	$\ell_2^{(1000)}$
100	5	2	23.95 (0.7)	1.86 (0.09)	0.42 (0.07)	0.57 (0.02)
100	5	3	21.49 (1.18)	2.05 (0.13)	0.31 (0.07)	0.59 (0.03)
100	10	2	26.14 (0.47)	1.79 (0.05)	0.44 (0.06)	0.54 (0.03)
100	10	3	19.53 (0.73)	1.91 (0.21)	0.43 (0.09)	0.6 (0.03)
300	5	2	35.18 (1.4)	3.54 (0.16)	0.24 (0.03)	0.53 (0.02)
300	5	3	32.10 (3.66)	3.28 (0.28)	0.21 (0.03)	0.47 (0.05)
300	10	2	35.46 (1.48)	3.40 (0.11)	0.31 (0.05)	0.52 (0.02)
300	10	3	39.49 (3.34)	5.34 (0.22)	0.29 (0.05)	0.51 (0.04)
500	5	2	43.20 (3.01)	5.29 (0.19)	0.20 (0.02)	0.48 (0.02)
500	5	3	46.93 (10.18)	5.30 (0.31)	0.20 (0.03)	0.36 (0.05)
500	10	2	58.33 (1.96)	4.78 (0.19)	0.25 (0.04)	0.48 (0.03)
500	10	3	60.42 (7.21)	6.96 (0.19)	0.26 (0.04)	0.42 (0.06)
1000	5	2	90.55 (19.42)	11.48 (0.74)	0.19 (0.01)	0.36 (0.03)
1000	5	3	125.23 (28.41)	10.91 (0.41)	0.19 (0.01)	0.24 (0.02)
1000	10	2	124.75 (12.86)	13.33 (0.72)	0.23 (0.02)	0.41 (0.02)
1000	10	3	103.95 (27.14)	14.29 (0.58)	0.24 (0.03)	0.3 (0.03)
2000	5	2	234.79 (71.89)	24.28 (1.42)	0.18 (0.01)	0.24 (0.02)
2000	5	3	239.01 (48.26)	28.42 (1.88)	0.17 (0.01)	0.19 (0.01)
2000	10	2	266.46 (60.08)	35.39 (2.26)	0.21 (0.02)	0.29 (0.03)
2000	10	3	266.57 (91.04)	35.37 (1.38)	0.22 (0.02)	0.24 (0.01)

Table A1. Average running time in second and normalized tensor mean squared error (with standard error in the parenthesis) for the 10000-iteration and 1000-iteration projected gradient descent algorithm over 100 replicated experiments.

Lastly, the cumulative regret and decision accuracy for model comparison are defined by:

$$\text{regret}(\widehat{D}_{\text{opt}}, D_{\text{opt}}) = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(P_{i,t,D_{i,\text{opt}}} - P_{i,t,\widehat{D}_{i,\text{opt}}}), \quad \text{acc}(\widehat{D}_{\text{opt}}, D_{\text{opt}}) = N^{-1}\sum_{i=1}^{N}\mathbf{1}(D_{i,\text{opt}} = \widehat{D}_{i,\text{opt}}).$$

A.2. Ablation analysis

In this section, we conduct the ablation studies to assess the effectiveness of our proposed framework. In particular, the mean squared error $\ell_2^{\text{group-w}}$ of the proposed framework is compared with other losses yielded by omitting three components individually:

1) The latent factor structures $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$: compare with the loss $\ell_2^{\text{GLM-w}}$ yielded by the logistic regression model stratified by the true membership of the treatments.

2) The grouping structure $\boldsymbol{M}$: compare with the loss $\ell_2^{\text{factor-w}}$ yielded by the latent factor model with the membership matrix $\boldsymbol{M}$ replacing by a latent factor matrix $\boldsymbol{U}_3$.

3) The inverse probability treatment weighting (IPTW): compare with the loss ℓ_2^{group} yielded by the unweighted minimization under the same latent factor block model.

Table A2 presents the ablation studies with the smallest normalized tensor mean squared error bolded, leading to the following conclusions: 1) The comparison of the proposed framework versus others highlights the advantages of leveraging latent factors across units and time for estimation; 2) The IPTW noticeably improves the results when the sample size N is small; 3) The hazard model adopting the grouping structure is particularly beneficial when the sample size is small. This is reasonable, as there may not be enough observations for certain treatments due to the limited sample size, and grouping treatments with homogeneous effects can enhance the estimation.

N	T	k	$\ell_2^{\text{GLM-w}}$	$\ell_2^{\text{factor-w}}$	ℓ_2^{group}	$\ell_2^{\text{group-w}}$
100	5	2	0.53 (0.02)	0.58 (0.02)	0.54 (0.03)	**0.42 (0.07)**
100	10	2	0.54 (0.03)	0.54 (0.03)	0.51 (0.03)	**0.44 (0.06)**
100	5	3	0.56 (0.02)	0.62 (0.02)	0.57 (0.03)	**0.31 (0.07)**
100	10	3	0.58 (0.03)	0.61 (0.03)	0.57 (0.03)	**0.43 (0.09)**
300	5	2	0.51 (0.01)	0.54 (0.02)	0.41 (0.03)	**0.24 (0.03)**
300	10	2	0.53 (0.02)	0.52 (0.02)	0.43 (0.03)	**0.31 (0.05)**
300	5	3	0.54 (0.01)	0.51 (0.05)	0.39 (0.03)	**0.21 (0.03)**
300	10	3	0.58 (0.01)	0.55 (0.04)	0.42 (0.04)	**0.29 (0.05)**
500	5	2	0.51 (0.01)	0.48 (0.03)	0.3 (0.03)	**0.20 (0.02)**
500	10	2	0.52 (0.01)	0.49 (0.03)	0.33 (0.04)	**0.25 (0.04)**
500	5	3	0.54 (0.01)	0.37 (0.06)	0.27 (0.03)	**0.20 (0.03)**
500	10	3	0.57 (0.01)	0.44 (0.07)	0.31 (0.03)	**0.26 (0.04)**
1000	5	2	0.51 (0.01)	0.31 (0.03)	0.21 (0.01)	**0.19 (0.01)**
1000	10	2	0.52 (0.01)	0.38 (0.04)	0.24 (0.01)	**0.23 (0.02)**
1000	5	3	0.54 (0.01)	**0.18 (0.02)**	0.2 (0.01)	0.19 (0.01)
1000	10	3	0.57 (0.01)	**0.23 (0.04)**	0.24 (0.01)	0.24 (0.03)
2000	5	2	0.5 (0.01)	**0.15 (0.02)**	0.18 (0.01)	0.18 (0.01)
2000	10	2	0.52 (0.01)	**0.19 (0.03)**	0.22 (0.01)	0.21 (0.02)
2000	5	3	0.53 (0.01)	**0.16 (0.02)**	0.18 (0.01)	0.17 (0.01)
2000	10	3	0.57 (0.01)	**0.20 (0.03)**	0.22 (0.01)	0.22 (0.02)

Table A2. Ablation analyses in terms of the normalized tensor mean squared error (with standard error in the parenthesis) over 100 replicated experiments

A.3. Additional real-data application

We provide an additional real-data application involving customer churn analysis of online retail to provide more empirical evidence. The data[2] are collected by an E-commerce company. Five important indicators for customer churn are considered: total number of registered devices (0 if less than 3, and 1 otherwise), the total number of used coupons in the last month (0 if less than the 50% quantile and 1 otherwise), the distance between warehouse to the customer (0 if less than the 50% quantile and 1 otherwise), whether the customer has complained or not, and post-complaint scores (0 for rates less than 3, and 1 otherwise). Next, we consider ten customer characteristics $\boldsymbol{X}_i$, including gender, marital status, city tier, preferred login device, preferred payment method, preferred order category, the total number of added addresses, percentage increases in orders from the last year, days since last order and the average cashback in the last month. We evaluate the models using the same cross-validation scheme as in Section 7, using the C-index and the average AUC to assess model performance. The best is **bolded**, and the second best is <u>underlined</u>. Our proposed framework continues to perform well on this E-commerce dataset based on these performance metrics.

B. Proofs

B.1. Assumptions

We first assume some regularity conditions to proceed with our illustrations of the theoretical results.

R1) (Positive Retention Probability) Let $p_{i,t,l} = P(\delta_i = 1, \mathcal{Y}_{i,t,l} = 1 \mid \mathcal{Y}_{i,t-1,l} = 1)$, we have $p_{\min} \leq p_{i,t,l}$ for any i, t and l;

R2) (Incoherent Tensor Parameter) Suppose $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ have orthonormal columns, there exists some constant μ_0 such that

$$\max\left\{\frac{N}{r_1}\|\boldsymbol{U}_1\|_{2,\infty}^2, \frac{T}{r_2}\|\boldsymbol{U}_2\|_{2,\infty}^2\right\} \leq \mu_0, \quad \max_{k\in\{1,2,3\}} \|\mathcal{M}_{(k)}(\mathcal{S})\| \leq \|\Theta\|_{\max}\sqrt{\frac{NTL}{\mu_0^{3/2}(r_1r_2r_3)^{1/2}}},$$

[2]https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

	C-index (↑)	average AUC (↑)
logit	0.33 ± 0.04	0.36 ± 0.03
RF	$\mathbf{0.42} \pm \mathbf{0.01}$	0.37 ± 0.02
NN	0.40 ± 0.00	0.41 ± 0.01
SVM	0.27 ± 0.03	$\underline{0.38 \pm 0.04}$
gradBoost	0.37 ± 0.01	0.25 ± 0.01
AdaBoost	0.27 ± 0.02	0.37 ± 0.01
Vote	0.38 ± 0.01	0.22 ± 0.01
Cox-PH	0.37 ± 0.01	0.37 ± 0.01
surv-RF	0.35 ± 0.01	0.33 ± 0.01
Cox-gradBoost	0.23 ± 0.01	0.18 ± 0.01
Cox-modelBoost	0.23 ± 0.01	0.26 ± 0.01
Proposed	$\underline{0.41 \pm 0.03}$	$\mathbf{0.44} \pm \mathbf{0.03}$

Table A3. Evaluation metrics on the E-commerce customer churn data.

where $\|\boldsymbol{U}\|_{2,\infty}^2 = \max_i \|\boldsymbol{U}_{i,:}\|^2$ and $\boldsymbol{U}_{i,:}$ is the i-th row vector of $\boldsymbol{U}$.

R3) (Non-degenerate Separation) The projected block means of Θ is defined as $\mathcal{S} = \Theta \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal \times_3 \boldsymbol{W}^\intercal$, where

$$\Delta_{\min}^2 = \Delta_{\min}(\mathcal{S})^2 = \min_{i_1 \neq i_2} \|\mathcal{M}_{(3)}(\mathcal{S})_{i_1,:} - \mathcal{M}_{(3)}(\mathcal{S})_{i_2,:}\|_2^2 > 0.$$

R4) (Balanced Clustering) There exists generic positive constants c and C such that

$$cL/r_3 \leq |\{l \in [L] : z_l = a\}| \leq CL/r_3, \quad \forall a = 1, \cdots, r_3,$$

where $|\cdot|$ represents the cardinality of a set.

Assumption R1) requires each unit to have a non-zero probability of maintaining the subscription at each time point t. This condition is necessary to establish the restricted strong convexity of the objective function $l(\Theta)$ in Section 5.1. Assumption R2) entails that the loading matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ should satisfy the incoherence condition, which is commonly imposed in the matrix/tensor completion literature (Candès & Recht, 2009; Ma et al., 2020; Cao et al., 2020; Cai et al., 2021). In particular, this condition indicates that each tensor entry contains a similar amount of information so that missing any of them will not prevent us from being able to recover the entire tensor. In addition, we also require an upper bound on the spectral norm of each matricization of the core tensor $\mathcal{S}$, which leads to an entry-wise upper bound on the absolute value of Θ together with the incoherence conditions. Assumption R3) requires that the projected mode-3 slices $\mathcal{M}_{(3)}(\mathcal{S})$ should have distinct rows; otherwise the number of the clustering size should be reduced to smaller. Assumption R4) is imposed to control the spectral norm of $\boldsymbol{M}$, and is widely used in mixture model clustering literature (Gao & Zhang, 2022; Han et al., 2022).

B.2. Proof of Theorem 5.1

The objective function for maximization is

$$\begin{aligned} l(\Theta) = &\sum_{i,t,l} w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l} = 1)\mathcal{Y}_{i,t,l} \log\{f(\theta_{i,t,l})\} \\ &+ \sum_{i,t,l} \delta_i w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l} = 1)(1 - \mathcal{Y}_{i,t,l}) \log\{1 - f(\theta_{i,t,l})\}, \end{aligned}$$

where the link function $f(\theta)$ is monotonically increasing. We further assume that $f(\theta)+f(-\theta)=1$, $f'(\theta)=f'(-\theta)$ and $f''(\theta)=-f''(-\theta)$. It follows from the expression of $l(\Theta)$ that

$$\begin{aligned}\frac{\partial l(\Theta)}{\partial \theta_{i,t,l}} &= w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)\mathcal{Y}_{i,t,l}\frac{f'(\theta_{i,t,l})}{f(\theta_{i,t,l})} - \delta_i w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)(1-\mathcal{Y}_{i,t,l})\frac{f'(\theta_{i,t,l})}{1-f(\theta_{i,t,l})},\\ \frac{\partial l^2(\Theta)}{\partial \theta_{i,t,l}^2} &= -w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)\mathcal{Y}_{i,t,l}\left[\frac{\{f'(\theta_{i,t,l})\}^2}{f^2(\theta_{i,t,l})} - \frac{f''(\theta_{i,t,l})}{f(\theta_{i,t,l})}\right]\\ &\quad - \delta_i w_i \mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)(1-\mathcal{Y}_{i,t,l})\left[\frac{f''(\theta_{i,t,l})}{1-f(\theta_{i,t,l})} + \frac{\{f'(\theta_{i,t,l})\}^2}{\{1-f(\theta_{i,t,l})\}^2}\right].\end{aligned}$$

Define

$$S(\Theta)=\left[\!\left[\frac{\partial l(\Theta)}{\partial \theta_{i,t,l}}\right]\!\right], \quad H(\Theta)=\left[\!\left[\frac{\partial l(\Theta)}{\partial \theta_{i,t,l}\partial \theta_{i',t',l'}}\right]\!\right],$$

where $S(\Theta)$ and $H(\Theta)$ are the collection of the first and second derivatives of $l(\Theta)$. By the second-order Taylor's Theorem, we expand $l(\widehat{\Theta})$ around the true parameter Θ and obtain

$$l(\widehat{\Theta}) = l(\Theta) + \langle S(\Theta), \widehat{\Theta}-\Theta^*\rangle + \frac{1}{2}\mathrm{vec}(\widehat{\Theta}-\Theta^*)^\intercal H(\tilde{\Theta})\mathrm{vec}(\widehat{\Theta}-\Theta^*), \tag{7}$$

where $\tilde{\Theta}=\gamma\Theta+(1-\gamma)\widehat{\Theta}$ for some $\gamma\in[0,1]$, and

$$\mathrm{vec}(\widehat{\Theta}-\Theta^*)^\intercal H(\tilde{\Theta})\mathrm{vec}(\widehat{\Theta}-\Theta^*) = \sum_{i,t,l}\left\{\frac{\partial l^2(\Theta)}{\partial \theta_{i,t,l}^2}\bigg|_{\Theta=\tilde{\Theta}}\right\}(\widehat{\theta}_{i,t,l}-\theta^*_{i,t,l})^2.$$

Let

$$L_\alpha = \sup_{|\theta|\le\alpha}\left[\frac{f'(\theta_{i,t,l})}{f(\theta_{i,t,l})}, \frac{f'(\theta_{i,t,l})}{1-f(\theta_{i,t,l})}\right], \quad \gamma_\alpha = \inf_{|\theta|\le\alpha}\left[\frac{\{f'(\theta_{i,t,l})\}^2}{f^2(\theta_{i,t,l})} - \frac{f''(\theta_{i,t,l})}{f(\theta_{i,t,l})}, \frac{f''(\theta_{i,t,l})}{1-f(\theta_{i,t,l})} + \frac{\{f'(\theta_{i,t,l})\}^2}{\{1-f(\theta_{i,t,l})\}^2}\right],$$

where $\alpha=\|\Theta\|_{\max}$ is the bound on the entry-wise magnitude of Θ. First, we bound the quadratic term in (7) as:

$$\sum_{i,t,l}\left\{\frac{\partial l^2(\Theta)}{\partial \theta_{i,t,l}^2}\bigg|_{\Theta=\tilde{\Theta}}\right\}(\widehat{\theta}_{i,t,l}-\theta^*_{i,t,l})^2 \le -\gamma_\alpha w_{\min}\sum_{i,t,l}\delta_i\mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta^*_{i,t,l})^2.$$

Lemma B.1. *Under the same conditions in Theorem 5.1 and* $\|\widehat{\Theta}-\Theta\|_F^2 \ge C_0\|\Theta\|_{\max}^2 T\log(N+T+K)/p_{\min}$*, there exists an constant* c_0*, such that, with probability greater than* $1-c_0(N+T+K)^{-2}$*,*

$$\sum_{i,t,l}\delta_i\mathbf{1}(\mathcal{Y}_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta^*_{i,t,l})^2 \ge \frac{p_{\min}}{2}\|\Delta_\Theta\|_F^2 - 2\vartheta,$$

where

$$\vartheta = \frac{C'' r_1 r_2 r_3 (N\vee T)}{\max(r_1,r_2,r_3)p_{\min}}\|\Theta\|_{\max}^2\log(N+T+K).$$

By Lemma B.1, we have

$$l(\widehat{\Theta}) \le l(\Theta) + \langle S(\Theta), \widehat{\Theta}-\Theta\rangle - \frac{\gamma_\alpha w_{\min}p_{\min}}{2}\|\widehat{\Theta}-\Theta\|_F^2 + 2\gamma_\alpha w_{\min}\vartheta,$$

where $\gamma_\alpha w_{\min}p_{\min}/2$ is curvature of $l(\Theta)$, and $\gamma_\alpha w_{\min}\vartheta$ is the tolerance term induced by the Rademacher complexity of $l(\Theta)$. Next, we bound the linear term in (7) as:

$$\begin{aligned}|\langle S(\Theta), \widehat{\Theta}-\Theta\rangle| &\le \|S(\Theta)\|\cdot\|\widehat{\Theta}-\Theta\|_*\\ &\le \|S(\Theta)\|\cdot 2\sqrt{\frac{r_1r_2r_3}{\max(r_1,r_2,r_3)}}\|\widehat{\Theta}-\Theta\|_F,\end{aligned}$$

where the last inequality is justified by Lemma 1, Wang & Li (2020).

Lemma B.2. *(Tensor Bernstein Inequality, Theorem 4.3, Luo et al. (2020)) Let $Z_1, \cdots, Z_N$ be independent tensor in $\mathbb{R}^{d\times d\times d}$, such that $\mathbb{E}(Z_i) = 0$ and $\|Z_i\| \le D_Z$ for all $i = 1, \cdots, N$. Let σ_Z^2 be such that*

$$\sigma_Z^2 \ge \max\left\{\left\|\mathbb{E}\left(\sum_{i=1}^N Z_i \overline{\square} \sum_{i=1}^N Z_i\right)\right\|, \left\|\mathbb{E}\left(\sum_{i=1}^N Z_i \underline{\square} \sum_{i=1}^N Z_i\right)\right\|\right\}.$$

Then for any $\alpha \ge 0$

$$\mathbb{P}\left(\left\|\sum_{i=1}^N Z_i\right\|^{\overline{\square}} \ge \alpha\right) \le d^2 \exp\left\{\frac{-\alpha^2}{2\sigma_Z^2 + (2D_Z\alpha)/3}\right\},$$

where $\overline{\square}$ and $\underline{\square}$ are two generalized Einstein products of tensors.

From Lemma B.2, we know

$$P(\|S(\Theta)\| \ge \alpha) \le (N+T+L)^2 \exp\left\{-\frac{\alpha^2}{2\sigma_S^2 + (2D_S\alpha)/3}\right\},$$

where

$$D_S = w_{\max} L_\alpha T^{1/2} \log(N+T+L), \quad \sigma_S^2 = w_{\max}^2 L_\alpha (N \vee T) \log(N+T+L).$$

Hence, implies that

$$\|S(\Theta)\| \le w_{\max}\sqrt{L_\alpha (N\vee T)} \log(N+T+L)$$

holds with probability greater than $1 - c_0(N+T+L)^{-2}$. Since $\widehat{\Theta}$ is the local maximizer, i.e., $\widehat{\Theta} = \arg\max_\Theta l(\Theta)$, we have

$$0 \le l(\widehat{\Theta}) - l(\Theta) \le \langle S(\Theta), \widehat{\Theta} - \Theta\rangle - \frac{\gamma_\alpha w_{\min} p_{\min}}{2}\|\widehat{\Theta} - \Theta\|_F^2 + 2\gamma_\alpha w_{\min}\vartheta,$$

which it gives us

$$\begin{aligned}
\|\widehat{\Theta} - \Theta\|_F^2 &\le \frac{C_1}{\gamma_\alpha w_{\min} p_{\min}}\|S(\Theta)\| \cdot \|\widehat{\Theta} - \Theta\|_* \\
&\quad + \frac{C_2 r_1 r_2 r_3 (N\vee T)}{\max(r_1,r_2,r_3) p_{\min}^2}\|\Theta\|_{\max}^2 \log(N+T+L) \\
&\le \frac{C_1}{\gamma_\alpha w_{\min} p_{\min}}\|S(\Theta)\| \cdot \sqrt{\frac{r_1 r_2 r_3}{\max(r_1,r_2,r_3)}}\|\widehat{\Theta} - \Theta\|_F \\
&\quad + \frac{C_2 r_1 r_2 r_3 (N\vee T)}{\max(r_1,r_2,r_3) p_{\min}^2}\|\Theta\|_{\max}^2 \log(N+T+L).
\end{aligned}$$

Therefore, with at least $1 - c_0(N+T+K)^{-2}$, we have $ab \le (a^2+b^2)/2$ and

$$\begin{aligned}
\|\widehat{\Theta} - \Theta\|_F^2 &\le \frac{C_1 w_{\max} L_\alpha}{\gamma_\alpha w_{\min} p_{\min}}\sqrt{\frac{r_1 r_2 r_3 (N\vee T)}{\max(r_1,r_2,r_3)}} \log(N+T+L) \cdot \|\widehat{\Theta} - \Theta\|_F \\
&\quad + \frac{C_2 r_1 r_2 r_3 (N\vee T)}{p_{\min}^2 \max(r_1,r_2,r_3)}\|\Theta\|_{\max}^2 \log(N+T+L) \\
&\le \frac{C_1' w_{\max}^2 L_\alpha^2}{\gamma_\alpha^2 w_{\min}^2 p_{\min}^2}\frac{r_1 r_2 r_3 (N\vee T)}{\max(r_1,r_2,r_3)} \log^2(N+T+L) + \frac{\|\widehat{\Theta} - \Theta\|_F^2}{2} \\
&\quad + \frac{C_2 r_1 r_2 r_3 (N\vee T)}{p_{\min}^2 \max(r_1,r_2,r_3)}\|\Theta\|_{\max}^2 \log(N+T+L).
\end{aligned}$$

To sum up, we conclude, with probability $1 - c_0(N+T+L)^{-2}$, we have

$$\begin{aligned}\|\widehat{\Theta} - \Theta\|_F^2 \le C_0 \|\Theta\|_{\max}^2 \frac{T}{p_{\min}} \log(N+T+L) \\ \vee C_1 \frac{L_\alpha^2}{\gamma_\alpha^2} \frac{w_{\max}^2}{w_{\min}^2} \frac{r_1 r_2 r_3 (N \vee T)}{p_{\min}^2 \max(r_1, r_2, r_3)} \log^2(N+T+L) \\ \vee C_2 \frac{r_1 r_2 r_3 (N \vee T)}{p_{\min}^2 \max(r_1, r_2, r_3)} \|\Theta\|_{\max}^2 \log(N+T+L).\end{aligned}$$

B.3. Proof of Theorem 5.3

The proof of Theorem 5.3 is divided into several steps to ease the understanding.

B.3.1. Step 1

We lay out some notations and technical Lemmas that will be useful for our main proof. First, we define the normalized membership matrices $\boldsymbol{W} = \boldsymbol{M}(\mathrm{diag}(\boldsymbol{1}_L^\intercal \boldsymbol{M}))^{-1}$. Next, we define the estimators of the projected block mean $\mathcal{S}$ and the projected mode-3 slices $\mathcal{F}$ by

$$\begin{aligned}&\mathcal{F} = \Theta \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal, \quad \mathcal{S} = \Theta \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal \times_3 \boldsymbol{W}^\intercal, \\ &\widehat{\mathcal{F}} = \widehat{\Theta} \times_1 (\widehat{\boldsymbol{U}}_1)^\intercal \times_2 (\widehat{\boldsymbol{U}}_2)^\intercal, \quad \widehat{\mathcal{S}} = \widehat{\Theta} \times_1 (\widehat{\boldsymbol{U}}_1)^\intercal \times_2 (\widehat{\boldsymbol{U}}_2)^\intercal \times_3 (\widehat{\boldsymbol{W}})^\intercal \\ &\tilde{\mathcal{F}} = \widehat{\Theta} \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal, \quad \tilde{\mathcal{S}} = \widehat{\Theta} \times_1 \boldsymbol{U}_1^\intercal \times_2 \boldsymbol{U}_2^\intercal \times_3 \boldsymbol{W}^\intercal.\end{aligned}$$

Next, we define the matricizations of each tensor $(\mathcal{S}, \mathcal{F})$ for the mode 3 by

$$\begin{aligned}&\boldsymbol{S} = \mathcal{M}_{(3)}(\mathcal{S}), \quad \widehat{\boldsymbol{S}} = \mathcal{M}_{(3)}(\widehat{\mathcal{S}}), \quad \tilde{\boldsymbol{S}} = \mathcal{M}_{(3)}(\tilde{\mathcal{S}}), \\ &\boldsymbol{F} = \mathcal{M}_{(3)}(\mathcal{F}) = \mathcal{M}_{(3)}(\Theta)(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2) = \mathcal{M}_{(3)}(\Theta)\boldsymbol{V}, \\ &\widehat{\boldsymbol{F}} = \mathcal{M}_{(3)}(\widehat{\mathcal{F}}) = \mathcal{M}_{(3)}(\widehat{\Theta})(\widehat{\boldsymbol{U}}_1 \otimes \widehat{\boldsymbol{U}}_2) = \mathcal{M}_{(3)}(\widehat{\Theta})\widehat{\boldsymbol{V}}, \\ &\tilde{\boldsymbol{F}} = \mathcal{M}_{(3)}(\tilde{\mathcal{F}}) = \mathcal{M}_{(3)}(\widehat{\Theta})(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2) = \mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V},\end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{U}_1 \otimes \boldsymbol{U}_2$ and $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{U}}_1 \otimes \widehat{\boldsymbol{U}}_2$.

Lemma B.3. *Under the same assumptions in Theorem 5.1, we have*

$$\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V}\| \lesssim \frac{r_3 g(\widehat{z}, z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z}, z) \|\widehat{\Theta} - \Theta\|_F}{\Delta_{\min}^2 \sqrt{NTL}}, \tag{8}$$

$$\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})\widehat{\boldsymbol{V}}\| \lesssim \frac{r_3 g(\widehat{z}, z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z}, z) \|\widehat{\Theta} - \Theta\|_F}{\Delta_{\min}^2 \sqrt{NTL}} + \frac{\kappa^2 r_3^{3/2} g(\widehat{z}, z) \|\widehat{\Theta} - \Theta\|_F}{\Delta_{\min}^2 \sqrt{L}}, \tag{9}$$

$$\|\widehat{\boldsymbol{W}}_{:,b}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})(\boldsymbol{V} - \widehat{\boldsymbol{V}})\| \lesssim \kappa^2 \sqrt{r_3/L} \|\widehat{\Theta} - \Theta\|_F, \tag{10}$$

$$g(\widehat{z}, z) \lesssim \Delta_{\min}^2 / r_3 \tag{11}$$

$$\sqrt{r_3/L} \lesssim \lambda_{r_3}(\boldsymbol{W}) \lesssim \|\boldsymbol{W}\| \lesssim \sqrt{r_3/L}, \tag{12}$$

$$\sqrt{L/r_3} \lesssim \lambda_{r_3}(\boldsymbol{M}) \lesssim \|\boldsymbol{M}\| \lesssim \sqrt{L/r_3}, \tag{13}$$

$$\|\widehat{\boldsymbol{V}} - \boldsymbol{V}\| \lesssim \frac{\kappa \|\widehat{\Theta} - \Theta\|_F}{\lambda_{\min}}, \tag{14}$$

$$\|\widehat{\boldsymbol{V}} - \boldsymbol{V}\|_{2,\max} \lesssim \frac{\kappa \|\widehat{\Theta} - \Theta\|_F}{\min(\delta_{r_1}, \delta_{r_2})}. \tag{15}$$

B.3.2. Step 2

In our nearest neighbor search algorithm, the estimated clustering label vector $\widehat{z}$ should satisfy:

$$\widehat{z}_l = \arg\min_{a \in [r_3]} \|\widehat{\boldsymbol{F}}_{l,:} - \widehat{\boldsymbol{S}}_{a,:}\|_2^2,$$

for $l = 1, \cdots, L$. Therefore, it is important to analyze the probably of the event:

$$\mathbf{1}(\widehat{z}_l = b) = \mathbf{1}\left(\widehat{z}_l = b, \|\widehat{\boldsymbol{F}}_{l,:} - \widehat{\boldsymbol{S}}_{b,:}\|_2^2 \leq \|\widehat{\boldsymbol{F}}_{l,:} - \widehat{\boldsymbol{S}}_{(z)_l,:}\|_2^2\right). \tag{16}$$

Assume $z_l = a$, we can check $\|\widehat{\boldsymbol{F}}_{l,:} - \widehat{\boldsymbol{S}}_{b,:}\|_2^2 \leq \|\widehat{\boldsymbol{F}}_{l,:} - \widehat{\boldsymbol{S}}_{a,:}\|_2^2$ is equivalent to

$$2\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle \leq -\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + F_l(a,b;\widehat{z}) + G_l(a,b;\widehat{z}) + H_l(a,b),$$

where

$$\begin{aligned}
\widehat{F}_l = F_l(a,b;\widehat{z}) &= 2\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\widehat{\boldsymbol{V}}, (\tilde{\boldsymbol{S}}_{a,:} - \widehat{\boldsymbol{S}}_{a,:}) - (\tilde{\boldsymbol{S}}_{b,:} - \widehat{\boldsymbol{S}}_{b,:})\rangle \\
&\quad + 2\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}(\boldsymbol{V} - \widehat{\boldsymbol{V}}), \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle, \\
\widehat{G}_l = G_l(a,b;\widehat{z}) &= \left(\|\Theta_{l:}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{a,:}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,a}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right) \\
&\quad - \left(\|\Theta_{l:}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{b,:}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right), \\
\widehat{H}_l = H_l(a,b) &= \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,a}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 \\
&\quad + \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2.
\end{aligned}$$

From these three error terms, the first two $F_l(a,b;\widehat{z})$ and $G_l(a,b;\widehat{z})$ are controlled by the differences between $(\widehat{\mathcal{S}}, \widehat{\mathcal{F}})$ and $(\tilde{\mathcal{S}}, \tilde{\mathcal{F}})$, the last one $H_l(a,b)$ is controlled by the differences between $(\tilde{\mathcal{S}}, \tilde{\mathcal{F}})$ and $(\mathcal{S}, \mathcal{F})$. If we ignore all these error terms, the event $2\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle \leq -\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2$ will constitute the oracle statistical loss after applying our algorithm if we are given the true label vector z. Then, we can show that

$$\begin{aligned}
(\widehat{z}_l = b) \subset &\left\{\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle \leq -\frac{1}{4}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2\right\} \\
&\bigcup \left\{\widehat{z}_l = b, \frac{1}{2}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 \leq \widehat{F}_l + \widehat{G}_l + \widehat{H}_l\right\}.
\end{aligned}$$

Recall the definition of $g(a,b)$, the misclassification loss for the estimated clustering label vector $\widehat{z}$ versus the truth will be:

$$\begin{aligned}
g(\widehat{z}, z) &= \min_{\pi \in \Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{\pi(\widehat{z})_{l,:}}\|_2^2 \\
&= \min_{\pi \in \Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3}\mathbf{1}\{\pi(\widehat{z})_l = b\}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2 \\
&\leq \min_{\pi \in \Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3}\{\mathbf{1}(\mathcal{E}_1) + \mathbf{1}(\mathcal{E}_2)\} \times \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2 \\
&= \widehat{g}_1 + \widehat{g}_2,
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{E}_1 &= \left\{\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle \leq -\frac{1}{4}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2\right\}, \\
\mathcal{E}_2 &= \left\{\widehat{z}_l = b, \frac{1}{2}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2 \leq \widehat{F}_l + \widehat{G}_l + \widehat{H}_l\right\}.
\end{aligned}$$

The following steps (Step 3 and Step 4) aim to bound $\widehat{g}_1$ and $\widehat{g}_2$, respectively.

B.3.3. STEP 3

In this step, we focus on proving the upper bound for $\widehat{g}_1 = \min_{\pi \in \Pi_{r_3}} \sum_{l=1}^{L}\sum_{b=1}^{r_3}\mathbf{1}(\mathcal{E}_1)\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2/L$. Recall that

$$\mathcal{E}_1 = \left\{\langle \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle \leq -\frac{1}{4}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2\right\},$$

we can decompose the probability of event $\mathcal{E}_1$ into three parts:

$$
\begin{aligned}
&P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\tilde{\boldsymbol{S}}_{b,:}\rangle \le -\frac{1}{4}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
\le& P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
&+P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\rangle \le -\frac{1}{16}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
&+P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \boldsymbol{S}_{b,:}-\tilde{\boldsymbol{S}}_{b,:}\rangle \le -\frac{1}{16}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right).
\end{aligned}
$$

Lemma B.4. *Under the same assumptions in Theorem 5.1, if the signal-to-noise ratio satisfies the condition in Theorem 5.3, there exist generic constants c_1 and c_2 such that*

$$
\begin{aligned}
P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right) &\le \frac{c_1}{(N+T+L)^2},\\
P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right) &\le \frac{c_1}{(N+T+L)^2}.
\end{aligned}
$$

With Lemma B.4, we can bound the expectation of $\widehat{g}_1$ if the signal-to-noise ratio satisfies the condition in Theorem 5.3:

$$
\begin{aligned}
\mathbb{E}\widehat{g}_1 &= \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b\in[r_3]/a} P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\tilde{\boldsymbol{S}}_{b,:}\rangle \le -\frac{1}{4}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\cdot\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2\\
&\lesssim \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b\in[r_3]/a} \exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}\cdot\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2.
\end{aligned}
$$

Since $\Delta_{\min}^2 \le \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2$ for any $a,b\in[r_3]$, it implies that

$$
\begin{aligned}
&\sum_{l=1}^{L}\sum_{b\in[r_3]/a}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2\cdot\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}\\
&\lesssim \{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}\sum_{l=1}^{L}\sum_{b\in[r_3]/a}\frac{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2}{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)}\cdot\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}\\
&\lesssim \{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}\sum_{l=1}^{L}\sum_{b\in[r_3]/a}\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}\\
&\lesssim \{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}L\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}.
\end{aligned}
$$

By Markov inequality, it yields that

$$
\begin{aligned}
&P\left[\widehat{g}_1 \le \mathbb{E}\widehat{g}_1\cdot\exp\left\{\frac{c_0}{2}\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}\right]\\
&\ge 1-\exp\left\{-\frac{c_0}{2}\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\}.
\end{aligned}
$$

Therefore, with probability at least $1-c(N+T+L)^{-2}$, we have

$$
\widehat{g}_1 \lesssim \{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T)r_1^2r_2^2r_3\mu_0^2\{\|\Theta\|_{\max}^2\vee(L_\alpha^2/\gamma_\alpha^2)\}}\right\},
$$

for some constant c.

B.3.4. STEP 4

This step aims to derive the upper bound for $\widehat{g}_2 = \min_{\pi\in\Pi_{r_3}} \sum_{l=1}^{L}\sum_{b=1}^{r_3} \mathbf{1}(\mathcal{E}_2)\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2/L$, i.e., the deterministic bounds for the error terms $\widehat{F}_l, \widehat{G}_l$ and $\widehat{H}_l$:

$$\widehat{g}_2 = \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3} \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2 \cdot \mathbf{1}\left\{\widehat{z}_l = b, \frac{1}{2}\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|_2^2 \le \widehat{F}_l + \widehat{G}_l + \widehat{H}_l\right\}.$$

Hereafter, we provide the deterministic upper bounds for $\widehat{F}_l$, $\widehat{G}_l$ and $\widehat{H}_l$, respectively.

(a) Upper bound for $F_l(a,b;\widehat{z})$.

$$\begin{aligned}
F_l(a,b;\widehat{z})^2 &\le 8|\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\widehat{\boldsymbol{V}}, (\tilde{\boldsymbol{S}}_{a,:} - \widehat{\boldsymbol{S}}_{a,:}) - (\tilde{\boldsymbol{S}}_{b,:} - \widehat{\boldsymbol{S}}_{b,:})\rangle|^2 \\
&+ 8|\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}}), \tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\rangle|^2 \\
&\le 32\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\widehat{\boldsymbol{V}}\|_{\max}^2 \cdot \max_{b\in[K]} \|\tilde{\boldsymbol{S}}_{b,:} - \widehat{\boldsymbol{S}}_{b,:}\|_F^2 \\
&+ 8\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|_{\max}^2 \cdot \|\tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\|^2 \\
&\le 64\left\{\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}\|_{\max}^2 + \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|_{\max}^2\right\} \cdot \max_{b\in[K]} \|\tilde{\boldsymbol{S}}_{b,:} - \widehat{\boldsymbol{S}}_{b,:}\|_2^2 \\
&+ 8\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|_{\max}^2 \cdot \|\tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\|^2,
\end{aligned} \tag{17}$$

where

$$\begin{aligned}
\|\tilde{\boldsymbol{S}}_{b,:} - \widehat{\boldsymbol{S}}_{b,:}\|_F^2 &= \|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V} + \widehat{\boldsymbol{W}}_{:,b}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|^2 \\
&\le 2\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V}\|^2 + 2\|\widehat{\boldsymbol{W}}_{:,b}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|^2 \\
&\overset{(8),(10)}{\lesssim} \left\{\frac{r_3 g(\widehat{z},z)}{\Delta_{\min}}\right\}^2 + \left\{\frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F}{\Delta_{\min}^2\sqrt{NTL}}\right\}^2 + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL} \\
&\lesssim r_3 g(\widehat{z},z) + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL},
\end{aligned} \tag{18}$$

and

$$\begin{aligned}
\|\tilde{\boldsymbol{S}}_{a,:} - \tilde{\boldsymbol{S}}_{b,:}\|^2 &= \|\tilde{\boldsymbol{S}}_{a,:} - \boldsymbol{S}_{a,:} + \boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:} + \boldsymbol{S}_{b,:} - \tilde{\boldsymbol{S}}_{b,:}\|^2 \\
&\le 3\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + 6\max_{a\in[r_3]} \|\tilde{\boldsymbol{S}}_{a,:} - \boldsymbol{S}_{a,:}\|^2 \\
&= 3\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + 6\max_{a\in[r_3]} \|\boldsymbol{W}_{:,a}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\boldsymbol{V}\|^2 \\
&\lesssim \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + \|\boldsymbol{W}_{:,a}\|^2 \cdot \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\boldsymbol{V}\|_{\max}^2 \\
&\overset{(12)}{\lesssim} \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + r_3/L \cdot \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F^2 \|\boldsymbol{V}\|_{2,\max}^2 \\
&\overset{(12)}{\lesssim} \|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2 + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL}.
\end{aligned} \tag{19}$$

Combing (17), (18) and (19), we obtain

$$\begin{aligned}
\frac{F_l(a,b;\widehat{z})^2}{\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2} &\lesssim \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}\|_{\max}^2 \cdot \left\{\frac{r_3 g(\widehat{z},z) + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL}}{\Delta_{\min}^2}\right\} \\
&+ \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|_{\max}^2 \cdot \left(1 + \frac{r_3 g(\widehat{z},z) + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL}}{\Delta_{\min}^2}\right),
\end{aligned}$$

which is bounded by

$$
\begin{aligned}
&\frac{F_l(a,b;\widehat{z})^2}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2} \\
\lesssim &\frac{1}{L}\sum_{l=1}^{L}\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}\|_{\max}^2\cdot\left\{\frac{r_3 g(\widehat{z},z)+\frac{\mu_0^2 r_1 r_2 r_3\|\widehat{\Theta}-\Theta\|_F^2}{NTL}}{\Delta_{\min}^2}\right\} \\
&+\frac{1}{L}\sum_{l=1}^{L}\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}(\boldsymbol{V}-\widehat{\boldsymbol{V}})\|_{\max}^2\cdot\left(1+\frac{r_3 g(\widehat{z},z)+\frac{\mu_0^2 r_1 r_2 r_3\|\widehat{\Theta}-\Theta\|_F^2}{NTL}}{\Delta_{\min}^2}\right) \\
\lesssim &\frac{\|\widehat{\Theta}-\Theta\|_F^2}{L}\cdot\|\boldsymbol{V}\|_{2,\max}^2\cdot\frac{r_3 g(\widehat{z},z)}{\Delta_{\min}^2}+\frac{\|\widehat{\Theta}-\Theta\|_F^2}{L}\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\|_{2,\max}^2\cdot\frac{r_3 g(\widehat{z},z)}{\Delta_{\min}^2} \\
\lesssim &\frac{\mu_0^2 r_1 r_2 r_3 g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL}. \qquad (20)
\end{aligned}
$$

(b) Upper bound for $G_l(a,b;\widehat{z})$.

$$
\begin{aligned}
G_l(a,b;\widehat{z}) = &\left(\|\Theta_{l:}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\|_F^2-\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right) \\
&-\left(\|\Theta_{l:}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\|_F^2-\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right) \\
= &\left(\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}+\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\|_F^2-\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right) \\
&-\left(\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}+\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\|_F^2-\|\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2\right) \\
= &\|\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\|_F^2-\|\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\|_F^2 \qquad (21) \\
&+2\langle\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\rangle \qquad (22) \\
&-2\langle\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\rangle, \qquad (23)
\end{aligned}
$$

where $\Theta_{l:}=\boldsymbol{W}_{:,(z)_l}^{\intercal}\Theta=\boldsymbol{W}_{:,a}^{\intercal}\{\widehat{\Theta}-(\widehat{\Theta}-\Theta)\}$. For the second part (22), we have

$$
\begin{aligned}
&\langle\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\rangle \\
=&\langle\boldsymbol{W}_{:,a}^{\intercal}\{\widehat{\Theta}-(\widehat{\Theta}-\Theta)\}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{a,:}\rangle \\
=&-\langle\boldsymbol{W}_{:,a}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle+\langle\boldsymbol{W}_{:,a}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},\widehat{\boldsymbol{S}}_{a,:}\rangle \\
=&-\langle\boldsymbol{W}_{:,a}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,a}-\widehat{\boldsymbol{W}}_{:,a})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle\rangle \\
&+\langle\boldsymbol{W}_{:,a}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},\widehat{\boldsymbol{S}}_{a,:}-\widehat{\boldsymbol{W}}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle \\
=&-\langle\boldsymbol{W}_{:,a}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,a}-\widehat{\boldsymbol{W}}_{:,a})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle\rangle,
\end{aligned}
$$

where $\widehat{\boldsymbol{S}}_{a,:}=\widehat{\boldsymbol{W}}_{:,a}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}$. For the third part (23), we have

$$
\begin{aligned}
&\langle\Theta_{l:}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\rangle \\
=&\langle\boldsymbol{W}_{:,a}^{\intercal}\{\widehat{\Theta}-(\widehat{\Theta}-\Theta)\}\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}-\widehat{\boldsymbol{S}}_{b,:}\rangle \\
=&\langle\boldsymbol{W}_{:,a}^{\intercal}\Theta\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\Theta\widehat{\boldsymbol{V}}+\boldsymbol{W}_{:,b}^{\intercal}\Theta\widehat{\boldsymbol{V}}-\boldsymbol{W}_{:,b}^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle \\
=&\langle(\boldsymbol{W}_{:,a}-\boldsymbol{W}_{:,b})^{\intercal}\Theta\widehat{\boldsymbol{V}}+\boldsymbol{W}_{:,b}^{\intercal}(\Theta-\widehat{\Theta})\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle \\
=&-\langle\boldsymbol{W}_{:,b}^{\intercal}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle \\
&+\langle(\boldsymbol{W}_{:,a}-\boldsymbol{W}_{:,b})^{\intercal}\Theta\widehat{\boldsymbol{V}},(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\widehat{\Theta}\widehat{\boldsymbol{V}}\rangle.
\end{aligned}
$$

Combined with (21), we further have

$$
\begin{aligned}
|G_l(a,b;\widehat{z})| \leq & \left| \|\boldsymbol{W}_{:,a}^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{a,:}\|_F^2 - \|\boldsymbol{W}_{:,b}^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{b,:}\|_F^2 \right| \\
& + 4 \max_{a\in[r_3]} \langle \boldsymbol{W}_{:,a}^\intercal (\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}, (\boldsymbol{W}_{:,a} - \widehat{\boldsymbol{W}}_{:,a})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} \rangle \\
& + 2 \left| \langle (\boldsymbol{W}_{:,a} - \boldsymbol{W}_{:,b})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}}, (\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} \rangle \right|.
\end{aligned}
$$

Then, we analyze these three terms separately. First, the first term is

$$
\begin{aligned}
& \left| \|\boldsymbol{W}_{:,a}^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{a,:}\|_F^2 - \|\boldsymbol{W}_{:,b}^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{b,:}\|_F^2 \right|^2 \\
\leq & \max_{a\in[r_3]} \|\boldsymbol{W}_{:,a}^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} - \widehat{\boldsymbol{S}}_{a,:}\|_F^4 \\
= & \max_{a\in[r_3]} \|(\boldsymbol{W}_{:,a} - \widehat{\boldsymbol{W}}_{:,a})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^4 \\
\overset{(8)}{\lesssim} & \frac{r_3^4 g^4(\widehat{z},z)}{\Delta_{\min}^4} + \frac{\mu_0^4 r_1^2 r_2^2 r_3^6 g^4(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^4}{\Delta_{\min}^8 N^2T^2L^2} + \frac{\kappa^8 r_3^6 g^4(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^4}{\Delta_{\min}^8 L^2} \\
\overset{(11)}{\lesssim} & \Delta_{\min}^4 + \frac{\mu_0^4 r_1^2 r_2^2 r_3^4 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^4}{\Delta_{\min}^4 N^2T^2L^2} + \frac{\kappa^8 r_3^4 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^4}{\Delta_{\min}^4 L^2}.
\end{aligned} \tag{24}
$$

The second term is:

$$
\begin{aligned}
& \max_a \left| \langle \boldsymbol{W}_{:,a}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}, (\boldsymbol{W}_{:,a} - \widehat{\boldsymbol{W}}_{:,a})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} \rangle \right|^2 \\
\leq & \max_a \|\boldsymbol{W}_{:,a}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|^2 \cdot \max_a \|(\boldsymbol{W}_{:,a} - \widehat{\boldsymbol{W}}_{:,a})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}}\|^2 \\
\leq & \max_a \|\boldsymbol{W}_{:,a}\|^2 \cdot \|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F^2 \cdot \|\boldsymbol{V}\|_{2,\max}^2 \\
& \times \left\{ \frac{r_3^2 g^2(\widehat{z},z)}{\Delta_{\min}^2} + \frac{\mu_0^2 r_1 r_2 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 NTL} + \frac{\kappa^4 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 L} \right\} \\
\lesssim & \frac{r_3}{L} \cdot \|\widehat{\Theta}-\Theta\|_F^2 \cdot \frac{\mu_0^2 r_1 r_2}{NT} \\
& \times \left\{ \Delta_{\min}^2 + \frac{\mu_0^2 r_1 r_2 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 NTL} + \frac{\kappa^4 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 L} \right\} \\
\lesssim & \frac{\mu_0^2 r_1 r_2 r_3 \Delta_{\min}^2 \|\widehat{\Theta}-\Theta\|_F^2}{NTL}.
\end{aligned} \tag{25}
$$

The last term is:

$$
\begin{aligned}
& \left| \langle (\boldsymbol{W}_{:,a} - \boldsymbol{W}_{:,b})^\intercal \Theta\widehat{\boldsymbol{V}}, (\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}} \rangle \right|^2 \\
\leq & \|(\boldsymbol{W}_{:,a} - \boldsymbol{W}_{:,b})^\intercal \Theta\widehat{\boldsymbol{V}}\|^2 \cdot \|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}}\|^2 \\
\leq & \|(\boldsymbol{S}_{:,a} - \boldsymbol{S}_{:,b})\boldsymbol{V}^\intercal \widehat{\boldsymbol{V}}\|^2 \cdot \|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \widehat{\Theta}\widehat{\boldsymbol{V}}\|^2 \\
\lesssim & \|\boldsymbol{S}_{:,a} - \boldsymbol{S}_{:,b}\|^2 \times \left\{ \Delta_{\min}^2 + \frac{\mu_0^2 r_1 r_2 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 NTL} \right\},
\end{aligned} \tag{26}
$$

where $\boldsymbol{W}_{:,a}^\intercal \Theta = \boldsymbol{S}_{a,:}\boldsymbol{V}^\intercal$. Combining (24), (25) and (26), we obtain

$$
\begin{aligned}
\frac{G_l(a,b;\widehat{z})^2}{\|\boldsymbol{S}_{a,:} - \boldsymbol{S}_{b,:}\|^2} & \lesssim \Delta_{\min}^2 + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL} + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL} \cdot \frac{\mu_0^2 r_1 r_2 r_3^3 g^2(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^4 NTL} \\
& \lesssim \Delta_{\min}^2 + \frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{NTL} \lesssim \Delta_{\min}^2,
\end{aligned} \tag{27}
$$

where $\|\widehat{\Theta}-\Theta\|_F^2 \lesssim NTL\Delta_{\min}^2/(\mu_0^2 r_1 r_2 r_3)$ holds with high probability by Theorem 5.1.

(c) Upper bound for $H_l(a,b)$.

$$
\begin{aligned}
\widehat{H}_l &= \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,a}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 + \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 \\
&= \|\boldsymbol{W}_{:,a}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 + \Big(\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big) \\
&- \Big(\|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big) \\
&= \|\boldsymbol{W}_{:,a}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 + \Big(\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big) \\
&- \Big(\|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\widehat{\Theta}\widehat{\boldsymbol{V}}\|_F^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big) \\
&= \Big(\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big) \\
&+ \|\boldsymbol{W}_{:,a}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 - \|\boldsymbol{W}_{:,b}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 \\
&+ 2\langle(\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:})\boldsymbol{V}^\intercal\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\rangle,
\end{aligned}
$$

where $\Theta_{l:} = \boldsymbol{W}_{:,(z)_l}^\intercal\Theta = \boldsymbol{W}_{:,a}^\intercal\{\widehat{\Theta}-(\widehat{\Theta}-\Theta)\}$ and $\boldsymbol{W}_{:,a}^\intercal\Theta = \boldsymbol{S}_{a,:}\boldsymbol{V}^\intercal$. For the first term, we have

$$
\begin{aligned}
&\Big|\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 - \|\Theta_{l:}\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal\Theta\widehat{\boldsymbol{V}}\|_F^2\Big| \\
=&\Big|\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2 - \|(\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:})\boldsymbol{V}^\intercal\widehat{\boldsymbol{V}}\|_F^2\Big| \\
\lesssim&\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2.
\end{aligned}
$$

And the second term is bounded by:

$$
\begin{aligned}
&\Big|\|\boldsymbol{W}_{:,a}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 - \|\boldsymbol{W}_{:,b}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2\Big| \\
\le&\max_a \|\boldsymbol{W}_{:,a}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\|_F^2 \\
\le&\max_a \|\boldsymbol{W}_{:,a}\|^2\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F^2\cdot\|\boldsymbol{V}\|_{2,\max}^2 \\
\lesssim&\frac{\mu_0^2 r_1 r_2 r_3\|\widehat{\Theta}-\Theta\|_F^2}{NTL}.
\end{aligned}
$$

The third term is

$$
\begin{aligned}
&\langle(\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:})\boldsymbol{V}^\intercal\widehat{\boldsymbol{V}} - \boldsymbol{W}_{:,b}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\rangle \\
\le&\|(\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:})\boldsymbol{V}^\intercal\widehat{\boldsymbol{V}}\|\cdot\|\boldsymbol{W}_{:,b}^\intercal(\widehat{\Theta}-\Theta)\widehat{\boldsymbol{V}}\| \\
\lesssim&\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|\cdot\frac{\mu_0^2 r_1 r_2 r_3\|\widehat{\Theta}-\Theta\|_F^2}{NTL}.
\end{aligned}
$$

Thus, we have

$$
\frac{|H_l(a,b)|}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2} \le c + \frac{\mu_0^2 r_1 r_2 r_3\|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL} \le \frac{1}{4}. \tag{28}
$$

Combining (20), (27) and (28), we obtain

$$
\begin{aligned}
\widehat{g}_2 &= \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3} \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2 \cdot \mathbf{1}\left\{\widehat{z}_l = b, \frac{1}{2}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2 \le \widehat{F}_l + \widehat{G}_l + \widehat{H}_l\right\} \\
&\le= \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3} \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2 \cdot \mathbf{1}\left\{\widehat{z}_l = b, \frac{1}{4}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2 \le \widehat{F}_l + \widehat{G}_l\right\} \\
&\le \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b=1}^{r_3} \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2 \cdot \mathbf{1}\left\{\widehat{z}_l = b, \|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^4 \le 64(\widehat{F}_l^2 + \widehat{G}_l^2)\right\} \\
&\le \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\sum_{b\in[r_3]/a} \mathbf{1}\left\{\widehat{z}_l = b\right\} \cdot 64\left(\frac{\widehat{F}_l^2}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2} + \frac{\widehat{G}_l^2}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2}\right) \\
&\le \min_{\pi\in\Pi_{r_3}} \frac{64}{L}\sum_{l=1}^{L}\max_{b\in[r_3]/a} \frac{\widehat{F}_l^2}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2} + \min_{\pi\in\Pi_{r_3}} \frac{64}{L}\sum_{l=1}^{L}\mathbf{1}\left\{\widehat{z}_l \ne z_l\right\}\max_{b\in[r_3]/a} \frac{\widehat{G}_l^2}{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|_2^2} \\
&\le \frac{g(\widehat{z},z)}{8}\frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL} + \min_{\pi\in\Pi_{r_3}} \frac{1}{L}\sum_{l=1}^{L}\mathbf{1}\left\{\widehat{z}_l \ne z_l\right\}\frac{\Delta_{\min}^2}{16} \\
&\le \frac{g(\widehat{z},z)}{8}\frac{\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL} + \frac{g(\widehat{z},z)}{16} \\
&\le \frac{g(\widehat{z},z)}{16}\left(1 + \frac{2\mu_0^2 r_1 r_2 r_3 \|\widehat{\Theta}-\Theta\|_F^2}{\Delta_{\min}^2 NTL}\right) \le \frac{g(\widehat{z},z)}{16}\left(1+\frac{1}{32}\right) \le \frac{g(\widehat{z},z)}{10}.
\end{aligned}
$$

Step 5 By combining Step 4 and Step3, we obtain

$$
\begin{aligned}
g(\widehat{z},z) &\le \widehat{g}_1 + \widehat{g}_2 \\
&\le \left\{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)\right\}\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T) r_1^2 r_2^2 r_3 \mu_0^2\left\{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)\right\}}\right\} + \frac{g(\widehat{z},z)}{10},
\end{aligned}
$$

which implies that

$$
g(\widehat{z},z) \lesssim \left\{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)\right\}\exp\left\{-\frac{w_{\min}^2 p_{\min}^2 NTL\max(r_1,r_2,r_3)\Delta_{\min}^2}{w_{\max}^2(N\vee T) r_1^2 r_2^2 r_3 \mu_0^2\left\{\|\Theta\|_{\max}^2 \vee (L_\alpha^2/\gamma_\alpha^2)\right\}}\right\},
$$

with probability at least $1 - c_0(N+T+L)^{-2}$.

B.4. Proof of Lemmas

B.4.1. Proof of Lemma B.1

We aim to prove that $\sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2$ is larger than $\|\Delta_\Theta\|_F^2 = \|\widehat{\Theta}-\Theta\|_F^2$ up to an additive term. Denote

$$
\begin{aligned}
&\|\Delta_\Theta\|_{\mathbb{E}}^2 = \mathbb{E}\left\{\sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right\} \ge p_{\min}\|\Delta_\Theta\|_F^2, \qquad (29)\\
&\|\Delta_\Theta\|_{\max}^2 = \max_{i,t,l}|(\Delta_\Theta)_{i,t,l}| \le 2|\Theta_{\max}|,
\end{aligned}
$$

where the expectation is taken with respect to $\mathbf{1}(Y_{i,t-1,l}=1)$. Given (29), we can show that

$$P\left\{\frac{p_{\min}}{2}\|\Delta_\Theta\|_F^2 \geq \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2+2\|\Delta_\Theta\|_{\max}^2\vartheta\right\}$$
$$\leq P\left\{\frac{\|\Delta_\Theta\|_{\mathbb{E}}^2}{2} \geq \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2+2\|\Delta_\Theta\|_{\max}^2\vartheta\right\}. \tag{30}$$

Therefore, our goal reduces to prove (30) is negligible.

Our proof for the restricted strong convexity proceeds by using the standard peeling argument. Let ξ be a constant larger than 1 (i.e., $\xi=2$) and define for every $\rho\geq 0$,

$$\mathcal{B}_{\rho\xi^{l-1}}=\left\{\Delta_\Theta\in\mathcal{B}:\rho\xi^{l-1}\leq\frac{\|\Delta_\Theta\|_{\mathbb{E}}^2}{\|\Theta\|_{\max}^2}\leq\rho\xi^l\right\},\quad l=1,2,\cdots,$$

and therefore $\mathcal{B}=\cup_{l=1}^\infty\mathcal{B}_{\rho\xi^{l-1}}$. For some $l\geq 1$ and $\Delta_\Theta\in\mathcal{B}_{\theta\xi^{l-1}}$, denote the event in (30) by

$$\mathcal{E}_l=\left\{\exists\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}:\left|\|\Delta_\Theta\|_{\mathbb{E}}^2-\sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right|\geq\frac{1}{2}\|\Delta_\Theta\|_{\mathbb{E}}^2+2\vartheta\geq\frac{1}{2\xi}\|\Theta\|_{\max}^2\rho\xi^l+2\vartheta\right\},$$

and $\mathcal{E}\subset\cup_{l=1}^\infty\mathcal{E}_l$. Let

$$\tilde{Z}_\rho=\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\left|\|\Delta_\Theta\|_{\mathbb{E}}^2-\sum_{i=1}^N\left\{\sum_{t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right\}\right|,$$

where $\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}$. Since each unit is at most being observed for T times for only one treatment, we know

$$\begin{aligned}\sigma_{\tilde{Z}_\rho}^2 &= \sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\sum_{i=1}^N\text{var}\left\{\sum_{t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right\}\\ &\leq \sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\sum_{i=1}^N\mathbb{E}\left\{\sum_{t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right\}^2\\ &\leq T\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\sum_{i=1}^N\mathbb{E}\left\{\sum_{t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right\}\\ &\leq T\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\|\Delta_\Theta\|_{\mathbb{E}}^2\leq T\|\Theta\|_{\max}^2\rho\xi^l,\end{aligned}$$

by the definition of $\mathcal{B}_{\rho\xi^{l-1}}$ which provides upper bounds for $\|\Delta_\Theta\|_{\mathbb{E}}^2$ and $\|\Delta_\Theta\|_{\max}^2$. By the symmetrization argument (Lemma 6.3, Ledoux & Talagrand (1991)),

$$\begin{aligned}\mathbb{E}(\tilde{Z}_\rho) &\leq 2\mathbb{E}\left\{\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\left|\sum_{i=1}^N\zeta_i\sum_{t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l}-\theta_{i,t,l})^2\right|\right\}\\ &\leq 4\sqrt{\frac{r_1r_2r_3\|\Theta\|_{\max}^2\rho\xi^l}{\max(r_1,r_2,r_3)p_{\min}}}\mathbb{E}\left\{\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\|\Delta_\Theta^{\text{Rad}}\|\right\}\\ &\leq C_1\|\Theta\|_{\max}^2\rho\xi^l+\frac{r_1r_2r_3}{\max(r_1,r_2,r_3)p_{\min}}\left[\mathbb{E}\left\{\sup_{\Delta_\Theta\in\mathcal{B}_{\rho\xi^{l-1}}}\|\Delta_\Theta^{\text{Rad}}\|\right\}\right]^2,\end{aligned}$$

where $\mathbb{E}\left\{\sup_{\Delta_\Theta \in \mathcal{B}_{\rho\xi^{l-1}}} \|\Delta_\Theta^{\text{Rad}}\|\right\}$ is the Rademacher complexity. Each entry of $\Delta_\Theta^{\text{Rad}}$ is defined by

$$(\Delta_\Theta^{\text{Rad}})_{i,t,l} = \zeta_i \mathbf{1}(Y_{i,t-1,l} = 1)(\widehat{\theta}_{i,t,l} - \theta_{i,t,l}) \cdot e_i(N) \otimes e_t(T) \otimes e_l(L),$$

where $e_i(N) \otimes e_t(T) \otimes e_l(K)$ is a zero tensor except its (i,t,l)-entry and $\{\zeta_i\}_{i=1}^N$ are i.i.d Rademacher random variables. Our next step is to obtain a close form of $\mathbb{E}\left\{\sup_{\Delta_\Theta \in \mathcal{B}_{\rho\xi^{l-1}}} \|\Delta_\Theta^{\text{Rad}}\|\right\}$ by tensor inequality. By Lemma B.2, we can show that

$$\begin{aligned} P\left(\sup_{\Delta_\Theta \in \mathcal{B}_{\rho\xi^{l-1}}} \|\Delta_\Theta^{\text{Rad}}\| \geq \alpha\right) &\leq (N+T+L)\exp\left[\frac{-\alpha^2}{2\sigma_{\Delta_\Theta}^2 + \{T^{1/2}\|\Theta\|_{\max}\log(N+T+L)\alpha\}/3}\right] \\ &\leq (N+T+L)\exp\left\{-\frac{3}{4}\frac{\alpha^2}{(N\vee T)\|\Theta\|_{\max}^2\log(N+T+L)}\right\}, \end{aligned}$$

where $\sigma_{\Delta_\Theta}^2 \lesssim (N\vee T)\|\Theta\|_{\max}^2\log(N+T+L)$. By Hölder's inequality, we have

$$\mathbb{E}\left\{\sup_{\Delta_\Theta \in \mathcal{B}_{\rho\xi^{l-1}}} \|\Delta_\Theta^{\text{Rad}}\|\right\} \lesssim \sqrt{(N\vee T)}\|\Theta\|_{\max}\{\log(N+T+L)\}^{1/2}.$$

Then, we invoke the Theorem 3 in Massart (2000) with $\varepsilon = 1$:

$$\begin{aligned} \mathbb{P}(\mathcal{E}_l) \leq& P\left\{\tilde{Z}_\rho \geq \frac{1}{2\xi}\|\Theta\|_{\max}^2\rho\xi^l + \frac{2r_1r_2r_3(N\vee T)}{\max(r_1,r_2,r_3)p_{\min}}\|\Theta\|_{\max}^2\log(N+T+L)\right\} \\ \leq& P\left\{\tilde{Z}_\rho \geq 2\sigma_{\tilde{Z}_\rho}\sqrt{x} + 34.5Tx + 2\mathbb{E}(\tilde{Z}_\rho)\right\} \\ \leq& \exp\left(-C_3x\right), \end{aligned}$$

where $x = C_3\rho\xi^l/T$. Therefore, we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_l) &= \mathbb{P}\left\{\sup_{\Delta_\Theta \in \mathcal{B}_{\rho\xi^{l-1}}}\left|\|\Delta_\Theta\|_{\mathbb{E}}^2 - \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l} - \theta_{i,t,l})^2\right| \geq \frac{1}{2\xi}\rho\xi^l + 2\vartheta\right\} \\ &\leq \exp\left(-\frac{C_3\rho\xi^l}{T}\right) \leq \exp\left(-\frac{C_3\rho l\log(\xi)}{T}\right) \leq \exp\left(-\frac{C_3'\rho l}{T}\right), \end{aligned}$$

since $\xi^l \geq l\log(\xi)$ for $\xi = 2$ and

$$\vartheta = \frac{r_1r_2r_3(N\vee T)}{\max(r_1,r_2,r_3)p_{\min}}\|\Theta\|_{\max}^2\log(N+T+L).$$

Thus,

$$\mathbb{P}(\mathcal{E}) \leq \sum_{l=1}^{\infty}\left\{\exp\left(-\frac{C_3'\rho}{T}\right)\right\}^l = \frac{\exp\left(-\frac{C_3'\rho}{T}\right)}{1-\exp\left(-\frac{C_3'\rho}{T}\right)} \leq 2\exp\left(-\frac{C_3'\rho}{T}\right),$$

and

$$\mathbb{P}\left\{\frac{1}{2}\|\Delta_\Theta\|_{\mathbb{E}}^2 \geq \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l} - \theta_{i,t,l})^2 + 2\vartheta\right\} \leq 2\exp\left(-\frac{C_3'\rho}{T}\right) \lesssim \frac{1}{(N+T+L)^2},$$

where $\rho = C'T\log(N+T+L)$ is determined to match the tail probability. To sum up, when $\|\widehat{\Theta} - \Theta\|_F^2 \geq C_0\|\Theta\|_{\max}^2 T\log(N+T+L)/p_{\min}$, we have

$$\begin{aligned} &\mathbb{P}\left\{\frac{p_{\min}}{2}\sum_{i,t,l}(\widehat{\theta}_{i,t,l} - \theta_{i,t,l})^2 \geq \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l} - \theta_{i,t,l})^2 + 2\vartheta\right\} \\ \leq& \mathbb{P}\left\{\frac{1}{2}\|\Delta_\Theta\|_{\mathbb{E}}^2 \geq \sum_{i,t,l}\mathbf{1}(Y_{i,t-1,l}=1)(\widehat{\theta}_{i,t,l} - \theta_{i,t,l})^2 + 2\vartheta\right\} \lesssim \frac{1}{(N+T+L)^2}, \end{aligned}$$

where

$$\vartheta = \frac{C'' r_1 r_2 r_3 (N \vee T)}{\max(r_1, r_2, r_3) p_{\min}} \|\Theta\|_{\max}^2 \log(N + T + L).$$

Thus, the proof of Lemma B.1 is completed.

B.4.2. PROOF OF LEMMA B.3

Proof. To prove (8), we notice that $\mathcal{M}_{(3)}(\Theta)\boldsymbol{V} = \boldsymbol{M}\mathcal{M}_{(3)}(\mathcal{S})$ and therefore

$$\begin{aligned}
&\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\Theta)\boldsymbol{V}\| \\
=&\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \boldsymbol{M}\boldsymbol{S}\| \\
=&\left\| \mathcal{M}_{(3)}(\mathcal{S})_{b,:} - \frac{\sum_{l=1}^L \mathcal{M}_{(3)}(\mathcal{S})_{(z)_l,:} \mathbf{1}(\widehat{z}_l = b)}{\sum_{l=1}^L \mathbf{1}(\widehat{z}_l = b)} \right\| \\
=&\left\| \frac{1}{\sum_{l=1}^L \mathbf{1}(\widehat{z}_l = b)} \sum_{l=1}^L \sum_{b' \in [r_3]/b} \mathbf{1}(z_l = b, \widehat{z}_l = b')\{\mathcal{M}_{(3)}(\mathcal{S})_{b,:} - \mathcal{M}_{(3)}(\mathcal{S})_{b',:}\} \right\| \\
\leq& C_0 \frac{r_3}{L\Delta_{\min}} \sum_{l=1}^L \sum_{b' \in [r_3]/b} \mathbf{1}(z_l = b, \widehat{z}_l = b') \|\mathcal{M}_{(3)}(\mathcal{S})_{b,:} - \mathcal{M}_{(3)}(\mathcal{S})_{b',:}\|^2 \\
\leq& C_0 \frac{r_3 g(\widehat{z}, z)}{\Delta_{\min}}.
\end{aligned}$$

Also, we know that

$$\begin{aligned}
\|\boldsymbol{W} - \widehat{\boldsymbol{W}}\| &\leq \{\lambda_{r_3}(\boldsymbol{M})\}^{-1} \cdot \|\boldsymbol{M}^\intercal \boldsymbol{W} - \boldsymbol{M}^\intercal \widehat{\boldsymbol{W}}\| \\
&\leq \{\lambda_{r_3}(\boldsymbol{M})\}^{-1} \cdot \|\boldsymbol{M}^\intercal \boldsymbol{W} - \boldsymbol{M}^\intercal \widehat{\boldsymbol{W}}\| \\
&\leq \{\lambda_{r_3}(\boldsymbol{M})\}^{-1} \cdot \|\boldsymbol{I} - \boldsymbol{M}^\intercal \widehat{\boldsymbol{W}}\|_F \\
&\overset{(13)}{\leq} \sqrt{\frac{r_3}{L}} \|\boldsymbol{I} - \boldsymbol{M}^\intercal \widehat{\boldsymbol{W}}\|_F,
\end{aligned} \tag{31}$$

where $\boldsymbol{M}^\intercal \boldsymbol{W} = \boldsymbol{I}$ by definition. For any $b \in [r_3]$, denote $\widehat{n}_b = \sum_{l=1}^L \mathbf{1}\{\widehat{z}_l = b\}$. For any $b \neq b'$, we have

$$(\boldsymbol{M}^\intercal \widehat{\boldsymbol{W}})_{b'b} = \frac{\sum_{l=1}^L \mathbf{1}\{z_l = b', \widehat{z}_l = b\}}{\widehat{n}_b}, \quad \delta_b = \sum_{b' \in [r_3]/b} (\boldsymbol{M}^\intercal \widehat{\boldsymbol{W}})_{b'b} = 1 - (\boldsymbol{M}^\intercal \widehat{\boldsymbol{W}})_{bb}.$$

Therefore,

$$
\begin{aligned}
\|\boldsymbol{I} - \boldsymbol{M}^\intercal \widehat{\boldsymbol{W}}\|_F &= \sqrt{\sum_{b\in[r_3]} \left\{ \delta_b^2 + \sum_{b'\in[r_3]/b} (\boldsymbol{M}^\intercal \widehat{\boldsymbol{W}})_{b'b}^2 \right\}} \\
&\leq \sqrt{\sum_{b\in[r_3]} \left\{ \delta_b^2 + \left(\sum_{b'\in[r_3]/b} (\boldsymbol{M}^\intercal \widehat{\boldsymbol{W}})_{b'b} \right)^2 \right\}} \\
&\leq \sqrt{2 \sum_{b\in[r_3]} \delta_b^2} \\
&\leq \sqrt{2} \sum_{b\in[r_3]} \delta_b \\
&\leq \sqrt{2} \sum_{b\in[r_3]} \frac{\sum_{l=1}^L \mathbf{1}\{z_l \neq b', \widehat{z}_l = b\}}{\widehat{n}_b} \\
&\leq \sqrt{2} \max_{b\in[r_3]} (\widehat{n}_b)^{-1} \sum_{l=1}^L \mathbf{1}(z_l \neq \widehat{z}_l) \\
&\lesssim \frac{r_3}{L} \cdot Lh(\widehat{z}, z) \lesssim \frac{r_3 g(\widehat{z}, z)}{\Delta_{\min}^2},
\end{aligned}
$$

where the last two inequalities are justified by (12). Then, we can show that

$$
\begin{aligned}
\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)\boldsymbol{V}\| &\leq \|\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b}\| \cdot \|\mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)\boldsymbol{V}\|_{\max} \\
&\overset{(31)}{\leq} C_1 \frac{r_3^{3/2} g(\widehat{z}, z)}{\Delta_{\min}^2 \sqrt{L}} \cdot \|\mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)\|_F \cdot \|\boldsymbol{V}\|_{2,\max} \\
&\leq C_1 \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z}, z) \|\widehat{\Theta} - \Theta\|_F}{\Delta_{\min}^2 \sqrt{NTL}}.
\end{aligned}
$$

By triangle inequality, it completes the proof for (8)

$$
\begin{aligned}
\|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V}\| &\leq \|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\Theta)\boldsymbol{V}\| \\
&\quad + \|(\boldsymbol{W}_{:,b} - \widehat{\boldsymbol{W}}_{:,b})^\intercal \mathcal{M}_{(3)}(\widehat{\Theta} - \Theta)\boldsymbol{V}\| \\
&\lesssim \frac{r_3 g(\widehat{z}, z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z}, z) \|\widehat{\Theta} - \Theta\|_F}{\Delta_{\min}^2 \sqrt{NTL}}.
\end{aligned}
$$

To prove (9),

$$
\begin{aligned}
&\|(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta})\widehat{\boldsymbol{V}}\| \\
&\le \|(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta})\boldsymbol{V}\| \\
&+ \|(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta})(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&\overset{(8)}{\lesssim} \frac{r_3 g(\widehat{z},z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F}{\Delta_{\min}^2\sqrt{NTL}} \\
&+ \|(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\mathcal{M}_{(3)}(\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&+ \|(\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b})^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&\lesssim \frac{r_3 g(\widehat{z},z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F}{\Delta_{\min}^2\sqrt{NTL}} \\
&+ \|\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b}\|\cdot\|\mathcal{M}_{(3)}(\Theta)\|\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\| \\
&+ \|\boldsymbol{W}_{:,b}-\widehat{\boldsymbol{W}}_{:,b}\|\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\|_{2,\max} \\
&\overset{(31),(14)}{\lesssim} \frac{r_3 g(\widehat{z},z)}{\Delta_{\min}} + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F}{\Delta_{\min}^2\sqrt{NTL}} + \frac{\kappa^2 r_3^{3/2} g(\widehat{z},z)\|\widehat{\Theta}-\Theta\|_F}{\Delta_{\min}^2\sqrt{L}}.
\end{aligned}
$$

To prove (10),

$$
\begin{aligned}
\|\widehat{\boldsymbol{W}}_{:,b}^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta})(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| &\le \|\widehat{\boldsymbol{W}}_{:,b}^{\intercal}\mathcal{M}_{(3)}(\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&+ \|\widehat{\boldsymbol{W}}_{:,b}^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&\le \|\boldsymbol{W}_{:,b}^{\intercal}\mathcal{M}_{(3)}(\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| + \|(\widehat{\boldsymbol{W}}_{:,b}^{\intercal}-\boldsymbol{W}_{:,b}^{\intercal})\mathcal{M}_{(3)}(\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&+ \|\boldsymbol{W}_{:,b}^{\intercal}\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| + \|(\widehat{\boldsymbol{W}}_{:,b}^{\intercal}-\boldsymbol{W}_{:,b}^{\intercal})\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)(\boldsymbol{V}-\widehat{\boldsymbol{V}})\| \\
&\lesssim \sqrt{r_3/L}\cdot\|\mathcal{M}_{(3)}(\Theta)\|\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\| + \frac{r_3^{3/2} g(\widehat{z},z)}{\Delta_{\min}^2\sqrt{L}}\cdot\|\mathcal{M}_{(3)}(\Theta)\|\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\| \\
&+ \sqrt{r_3/L}\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\|_{2,\max} \\
&+ \frac{r_3^{3/2} g(\widehat{z},z)}{\Delta_{\min}^2\sqrt{L}}\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F\cdot\|\boldsymbol{V}-\widehat{\boldsymbol{V}}\|_{2,\max} \\
&\overset{(14)}{\lesssim} \sqrt{r_3/L}\cdot\kappa^2\|\widehat{\Theta}-\Theta\|_F + \frac{r_3^{3/2} g(\widehat{z},z)}{\Delta_{\min}^2\sqrt{L}}\cdot\kappa^2\|\widehat{\Theta}-\Theta\|_F \\
&+ \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{1/2}}{\sqrt{NTL}}\|\widehat{\Theta}-\Theta\|_F + \frac{\mu_0 r_1^{1/2} r_2^{1/2} r_3^{3/2} g(\widehat{z},z)}{\Delta_{\min}^2\sqrt{NTL}}\|\widehat{\Theta}-\Theta\|_F \\
&\lesssim \kappa^2\sqrt{r_3/L}\|\widehat{\Theta}-\Theta\|_F.
\end{aligned}
$$

To prove (14), we have

$$
\begin{aligned}
\|\widehat{\boldsymbol{V}}-\boldsymbol{V}\| &= \|\widehat{\boldsymbol{U}}_1\otimes\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_1\otimes\boldsymbol{U}_2\| \\
&= \|\widehat{\boldsymbol{U}}_1\otimes(\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_2)+(\widehat{\boldsymbol{U}}_1-\boldsymbol{U}_1)\otimes\boldsymbol{U}_2\| \\
&\le \|\widehat{\boldsymbol{U}}_1\otimes(\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_2)\| + \|(\widehat{\boldsymbol{U}}_1-\boldsymbol{U}_1)\otimes\boldsymbol{U}_2\| \\
&\lesssim \left\{\|\widehat{\boldsymbol{U}}_1-\boldsymbol{U}_1\| + \|\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_2\|\right\} \overset{\text{Lemma } B.5}{\lesssim} \frac{\kappa\|\widehat{\Theta}-\Theta\|_F}{\lambda_{\min}}.
\end{aligned}
$$

Since $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are only identifiable up to rotation and permutation, we instead focus on the upper bound of $\min_{\boldsymbol{R}_1\in\mathbb{O}_{r_1}}\|\widehat{\boldsymbol{U}}_1-\boldsymbol{U}_1\boldsymbol{R}_1\|$ and $\min_{\boldsymbol{R}_2\in\mathbb{O}_{r_2}}\|\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_2\boldsymbol{R}_2\|$, where $\mathbb{O}_r$ is the collection of all r-by-r matrices with orthonormal columns. Lemma B.5 is a variant of the Davis-Kahan $\sin(\Theta)$ Theorem from Yu et al. (2015), which bound the distance between subspaces spanned by the population eigenvectors and their sample versions.

For (14) and (15), the proof is similar to Theorem 4, Yu et al. (2015):

$$\begin{aligned}
\|\boldsymbol{V}\|_{2,\max}^2 &= \max_j \|e_j^\intercal(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2)\|_2^2 \\
&\le \max_{i,t} \|e_i^\intercal \boldsymbol{U}_1\|^2 \cdot \|e_t^\intercal \boldsymbol{U}_2\|^2 \text{ (Cauchy-Schwatz inequality)} \\
&\le \|\boldsymbol{U}_1\|_{2,\max}^2 \|\boldsymbol{U}_2\|_{2,\max}^2 \\
&\le \frac{\mu_0^2 r_1 r_2}{NT},
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{V} - \widehat{\boldsymbol{V}}\|_{2,\max} &= \max_j \|e_j^\intercal(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2 - \widehat{\boldsymbol{U}}_1 \otimes \widehat{\boldsymbol{U}}_2)\| \\
&= \max_j \|e_j^\intercal(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2 - \boldsymbol{U}_1 \otimes \widehat{\boldsymbol{U}}_2 + \boldsymbol{U}_1 \otimes \widehat{\boldsymbol{U}}_2 - \widehat{\boldsymbol{U}}_1 \otimes \widehat{\boldsymbol{U}}_2)\| \\
&\le \max_j \|e_j^\intercal(\boldsymbol{U}_1 \otimes \boldsymbol{U}_2 - \boldsymbol{U}_1 \otimes \widehat{\boldsymbol{U}}_2)\| + \max_{j'} \|e_{j'}^\intercal(\boldsymbol{U}_1 \otimes \widehat{\boldsymbol{U}}_2 - \widehat{\boldsymbol{U}}_1 \otimes \widehat{\boldsymbol{U}}_2)\| \\
&\le \|\boldsymbol{U}_1\|_{2,\max} \|\widehat{\boldsymbol{U}}_2 - \boldsymbol{U}_2\|_{2,\max} + \|\widehat{\boldsymbol{U}}_2\|_{2,\max} \|\widehat{\boldsymbol{U}}_1 - \boldsymbol{U}_1\|_{2,\max} \\
&\le \sqrt{\frac{\mu_0 r_1}{N}} \|\widehat{\boldsymbol{U}}_2 - \boldsymbol{U}_2\|_{2,\max} + \sqrt{\frac{\mu_0 r_2}{T}} \|\widehat{\boldsymbol{U}}_1 - \boldsymbol{U}_1\|_{2,\max}.
\end{aligned}$$

Under the assumption of eigengap, we are able to bound the perturbation of individual eigenvectors:

$$\|\widehat{u}_{2,i} - u_{2,i}\|_2 \lesssim \frac{\lambda_{\max} \|\widehat{\Theta} - \Theta\|_F}{\min(\lambda_{i-1}^2 - \lambda_i^2, \lambda_i^2 - \lambda_{i+1}^2)}.$$

Assume for any $i \in [r_2]$, the interval $[\lambda_i - \delta_{r_2}, \lambda_i + \delta_{r_2}]$ does not contain any eigenvalues of $\mathcal{M}_{(2)}(\Theta)$ other than λ_i:

$$\|\widehat{u}_{2,i} - u_{2,i}\|_2 \lesssim \frac{\lambda_{\max} \|\widehat{\Theta} - \Theta\|_F}{\min(\lambda_{i-1}^2 - \lambda_i^2, \lambda_i^2 - \lambda_{i+1}^2)} \lesssim \frac{\kappa \|\widehat{\Theta} - \Theta\|_F}{\delta_{r_2}},$$

which implies $\|\widehat{\boldsymbol{U}}_2 - \boldsymbol{U}_2\|_{2,\max} \le \kappa \|\widehat{\Theta} - \Theta\|_F / \delta_{r_2}$. Identical bounds also hold if $\widehat{u}_{2,i}$ and $u_{2,i}$ are replaced with $\widehat{u}_{1,i}$ and $u_{1,i}$. Suppose

$$\delta_{r_1} \ge \sqrt{\frac{N}{\mu_0 r_1}} \cdot \kappa \|\widehat{\Theta} - \Theta\|_F, \quad \delta_{r_2} \ge \sqrt{\frac{T}{\mu_0 r_2}} \cdot \kappa \|\widehat{\Theta} - \Theta\|_F,$$

we claim that $\|\boldsymbol{V} - \widehat{\boldsymbol{V}}\|_{2,\max}^2 \le \|\boldsymbol{V}\|_{2,\max}^2$. □

B.4.3. Proof of Lemma B.4

First, we want to prove the probability $P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)$ is negligible. We can show that

$$
\begin{aligned}
&P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
&\le P(-\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|(\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:})\boldsymbol{V}^\intercal\|_{\max} \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2)\\
&\le P(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|\cdot\|\boldsymbol{V}\|_{2,\max} \ge \frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2)\\
&\le P(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|\boldsymbol{U}_1\|_{2,\max}\|\boldsymbol{U}_2\|_{2,\max} \ge \frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|)\\
&\le P\left(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|^2 \ge \frac{1}{64}\frac{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2}{\|\boldsymbol{U}_1\|_{2,\max}^2\|\boldsymbol{U}_2\|_{2,\max}^2}\right)\\
&\le P\left(\sum_{l=1}^{L}\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|^2 \ge \frac{1}{64}\frac{L\Delta_{\min}^2}{\|\boldsymbol{U}_1\|_{2,\max}^2\|\boldsymbol{U}_2\|_{2,\max}^2}\right)\\
&\le P\left(\|\widehat{\Theta}-\Theta\|_F^2 \ge \frac{1}{64}\frac{NTL\Delta_{\min}^2}{\mu_0^2 r_1 r_2}\right),
\end{aligned}
$$

where under Assumption R2):

$$
\|\boldsymbol{U}_1\|_{2,\max}^2 \le \frac{\mu_0 r_1}{N}, \quad \|\boldsymbol{U}_2\|_{2,\max}^2 \le \frac{\mu_0 r_2}{T}.
$$

To ensure the event $\|\widehat{\Theta}-\Theta\|_F^2 \le NTL\Delta_{\min}^2/(64\mu_0^2 r_1 r_2)$ holds with high probability, we need to have

$$
\frac{NTL\Delta_{\min}^2}{\mu_0^2 r_1 r_2} \ge \frac{c_0(N\vee T)r_1 r_2 r_3 \log(N+T+L)\|\Theta\|_{\max}^2}{p_{\min}^2 \max(r_1,r_2,r_3)}, \tag{32}
$$

$$
\frac{NTL\Delta_{\min}^2}{\mu_0^2 r_1 r_2} \ge \frac{c_0 w_{\max}^2(N\vee T)r_1 r_2 r_3 \log^2(N+T+L)(L_\alpha^2/\gamma_\alpha^2)}{w_{\min}^2 p_{\min}^2 \gamma_\alpha^2 \max(r_1,r_2,r_3)}, \tag{33}
$$

under the estimation error in Theorem 5.1. Thus, we claim that

$$
P\left(\|\widehat{\Theta}-\Theta\|_F^2 \ge \frac{1}{64}\frac{NTL\Delta_{\min}^2}{\mu_0^2 r_1 r_2}\right) \le \frac{c}{(N+T+L)^2},
$$

under the conditions (32) and (33). Next, we show that $P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)$ is negligible

$$
\begin{aligned}
&P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\rangle \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
&\le P(-\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|(\tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:})\boldsymbol{V}^\intercal\|_{\max} \le -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2)\\
&\le P(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|\tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\|\cdot\|\boldsymbol{V}\|_{2,\max} \ge \frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2)\\
&\le P(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\cdot\|\tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\|\cdot\|\boldsymbol{U}_1\|_{2,\max}\|\boldsymbol{U}_2\|_{2,\max} \ge \frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|).
\end{aligned} \tag{34}
$$

One can observe that

$$
\begin{aligned}
\|\tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\| &= \|\boldsymbol{M}_{:,a}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\boldsymbol{V}\|\\
&\le \|\boldsymbol{M}_{:,a}^\intercal \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|\cdot\|\boldsymbol{V}\|_{\max}\\
&\le \|\boldsymbol{M}_{:,a}^\intercal\|\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F\cdot\max_{i,j}\|e_i^\intercal \boldsymbol{V} e_j\|\\
&\le \sqrt{\frac{r_3}{L}}\|\boldsymbol{U}_1\|_{2,\max}\|\boldsymbol{U}_2\|_{2,\max}\cdot\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F.
\end{aligned}
$$

Plug it back to (34), we obtain

$$
\begin{aligned}
&P\left(\langle \mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\boldsymbol{V}, \tilde{\boldsymbol{S}}_{a,:}-\boldsymbol{S}_{a,:}\rangle \leq -\frac{1}{8}\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2\right)\\
&\leq P\left(\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F \geq \frac{\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2}{8\sqrt{\frac{r_3}{L}}\|\boldsymbol{U}_1\|_{2,\max}^2\|\boldsymbol{U}_2\|_{2,\max}^2}\right)\\
&\leq P\left(\sum_{l=1}^{L}\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)_{l,:}\|\|\mathcal{M}_{(3)}(\widehat{\Theta}-\Theta)\|_F \geq \frac{L\|\boldsymbol{S}_{a,:}-\boldsymbol{S}_{b,:}\|^2}{8\sqrt{\frac{r_3}{L}}\|\boldsymbol{U}_1\|_{2,\max}^2\|\boldsymbol{U}_2\|_{2,\max}^2}\right)\\
&\leq P\left(\|\widehat{\Theta}-\Theta\|_F^2 \geq \frac{NTL^{3/2}\Delta_{\min}^2}{8\mu_0^2 r_1 r_2\sqrt{r_3}}\right),
\end{aligned}
$$

which is negligible implied by the signal-to-noise ratio condition since $L \geq r_3$:

$$
\frac{NTL^{3/2}\Delta_{\min}^2}{\mu_0^2 r_1 r_2\sqrt{r_3}} \geq \frac{NTL\Delta_{\min}^2}{\mu_0^2 r_1 r_2} \geq \frac{c_0 L_\alpha^2 (N\vee T) r_1 r_2 r_3 (\|\Theta\|_{\max}^2 \vee \sigma^2)\log^2(N+T+L)}{p_{\min}^4\gamma_\alpha^2\max(r_1,r_2,r_3)}.
$$

Thus, which completes our proof of Lemma B.4.

Lemma B.5. *Let*

$$
\begin{aligned}
\lambda_{\max} &= \max\{\|\mathcal{M}_{(1)}(\mathcal{Y})\|, \|\mathcal{M}_{(2)}(\mathcal{Y})\|, \|\mathcal{M}_{(3)}(\mathcal{Y})\|\},\\
\lambda_{\min} &= \min\left[\lambda_{r_1}\{\mathcal{M}_{(1)}(\mathcal{Y})\}, \lambda_{r_2}\{\mathcal{M}_{(2)}(\mathcal{Y})\}, \lambda_{r_3}\{\mathcal{M}_{(3)}(\mathcal{Y})\}\right],
\end{aligned}
$$

and $\kappa = \lambda_{\max}/\lambda_{\min}$, then we have:

$$
\min_{\boldsymbol{R}_1\in\mathbb{O}_{r_1}}\|\widehat{\boldsymbol{U}}_1-\boldsymbol{U}_1\boldsymbol{R}_1\|_F \leq C\frac{\kappa\|\widehat{\Theta}-\Theta\|_F}{\lambda_{\min}}, \quad \min_{\boldsymbol{R}_2\in\mathbb{O}_{r_2}}\|\widehat{\boldsymbol{U}}_2-\boldsymbol{U}_2\boldsymbol{R}_2\| \leq C\frac{\kappa\|\widehat{\Theta}-\Theta\|_F}{\lambda_{\min}}.
$$

The proof of Lemma B.5 is provided in Yu et al. (2015), Theorem 4.

Lemma B.6. *Under condition (11), we have $r_3/L \lesssim |l\in[L]:(\widehat{z})_l = a| \lesssim r_3/L$ for any $a\in[r_3]$. Moreover,*

$$
\begin{aligned}
\sqrt{L/r_3} \lesssim \lambda_{r_3}(\boldsymbol{M}) \leq \|\boldsymbol{M}\| \lesssim \sqrt{L/r_3},\\
\sqrt{r_3/L} \lesssim \lambda_{r_3}(\boldsymbol{W}) \leq \|\boldsymbol{W}\| \lesssim \sqrt{r_3/L}.
\end{aligned}
$$

The above inequalities also hold by replacing $\boldsymbol{M}$, $\boldsymbol{W}$ with $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{W}}$, respectively.

The proof of Lemma B.6 is provided in Han et al. (2022), Lemma 4.