Journal of Machine Learning Research 22 (2021) 1-43 Submitted 9/19; Revised 1/21; Published 1/21

Sparse Tensor Additive Regression

Botao Hao HAOBOTAO000@GMAIL.COM
Deepmind
5 New Street, London, UK

Boxiang Wang BOXIANG-WANG@UIOWA.EDU
Department of Statistics and Actuarial Science
The University of Iowa
Iowa City, IA 52242, USA

Pengyuan Wang PENGYUAN@UGA.EDU
Department of Marketing
University of Georgia
Athens, GA 30602, USA

Jingfei Zhang EZHANG@BUS.MIAMI.EDU
Department of Management Science
University of Miami
Coral Gables, FL 33146, USA

Jian Yang JIANYANG@VERIZONMEDIA.COM
Yahoo Research
Verizon Media
Sunnyvale, CA 94089, USA

Will Wei Sun SUN244@PURDUE.EDU
Krannert School of Management
Purdue University
West Lafayette, IN 47907, USA

Editor: Francis Bach

Abstract

Tensors are becoming prevalent in modern applications such as medical imaging and digital marketing. In this paper, we propose a sparse tensor additive regression (STAR) that models a scalar response as a flexible nonparametric function of tensor covariates. The proposed model effectively exploits the sparse and low-rank structures in the tensor additive regression. We formulate the parameter estimation as a non-convex optimization problem, and propose an efficient penalized alternating minimization algorithm. We establish a non-asymptotic error bound for the estimator obtained from each iteration of the proposed algorithm, which reveals an interplay between the optimization error and the statistical rate of convergence. We demonstrate the efficacy of STAR through extensive comparative simulation studies, and an application to the click-through-rate prediction in online advertising.

Keywords: additive models; low-rank tensor; non-asymptotic analysis; non-convex optimization; tensor regression.

1. Introduction

Tensor data have recently become popular in a wide range of applications such as medical imaging (Zhou et al., 2013; Li and Zhang, 2017; Sun and Li, 2017), digital marketing (Zhe et al., 2016; Sun et al., 2017), video processing (Guo et al., 2012), and social network analysis (Park and Chu, 2009; Hoff, 2015), among many others. In such applications, a fundamental statistical tool is *tensor regression*, a modern high-dimensional regression method that relates a scalar response to tensor covariates. For example, in neuroimaging analysis, an important objective is to predict clinical outcomes using subjects' brain imaging data. This can be formulated as a tensor regression problem by treating the clinical outcomes as the response and the brain images as the tensor covariates. Another example is in the study of how advertisement placement affect users' clicking behavior in online advertising. This again can be formulated as a tensor regression problem by treating the daily overall click-through rate (CTR) as the response and the tensor that summarizes the impressions (i.e., view counts) of different advertisements on different devices (e.g., phone, computer, etc.) as the covariate. In Section 6, we consider such an online advertising application.

Denote y_i as a scalar response and $\mathcal{X}_i \in \mathbb{R}^{p_1 \times p_2 \cdots \times p_m}$ as an m-way tensor covariate, $i = 1, 2, \ldots, n$. A general tensor regression model can be formulated as

$$y_i = \mathcal{T}^*(\mathcal{X}_i) + \epsilon_i, \; i = 1, 2, \ldots, n,$$

where $\mathcal{T}^*(\cdot) : \mathbb{R}^{p_1 \times p_2 \cdots \times p_m} \to \mathbb{R}$ is an unknown regression function, $\{\epsilon_i\}_{i=1}^n$ are scalar observation noises. Many existing methods assumed a linear relationship between the response and the tensor covariates by considering $\mathcal{T}^*(\mathcal{X}_i) = \langle \mathcal{B}, \mathcal{X}_i \rangle$ for some low-rank tensor coefficient $\mathcal{B}$ (Zhou et al., 2013; Rabusseau and Kadri, 2016; Yu and Liu, 2016; Guhaniyogi et al., 2017; Raskutti et al., 2019). In spite of its simplicity, the linear assumption could be restrictive and difficult to satisfy in real applications. Consider the online advertising data in Section 6 as an example. Figure 1 shows the marginal relationship between the overall CTR and the impressions of an advertisement delivered on phone, tablet, and PC, respectively. It is clear that the relationship between the response (i.e., the overall CTR) and the covariate (i.e., impressions across three devices) departs notably from the linearity assumption. A few work considered more flexible tensor regressions by treating $\mathcal{T}^*(\cdot)$ as a nonparametric function (Suzuki et al., 2016; Kanagawa et al., 2016). In particular, Suzuki et al. (2016) proposed a general nonlinear model where the true function $\mathcal{T}^*(\cdot)$ is consisted of components from a reproducing kernel Hilbert space, and used an alternating minimization estimation procedure; Kanagawa et al. (2016) considered a Bayesian approach that employed a Gaussian process prior in learning the nonparametric function $\mathcal{T}^*(\cdot)$ on the reproducing kernel Hilbert space. One serious limitation of both work is that they assume that the tensor covariates are *exact low-rank*. This assumption is difficult to satisfy in practice, as most tensor covariates are not exact low-rank. When the tensor covariates are not exact low-rank, the performance of these two methods deteriorates dramatically; see Section 5.2 for more details. In addition, the Gaussian process approach is computationally very expensive, which severely limits its application in problems with high-dimensional tensor covariates.

In this paper, we develop a flexible and computationally feasible tensor regression framework, which accommodates the nonlinear relationship between the response and the tensor covariate, and is highly interpretable. Specifically, for an m-way tensor covariate

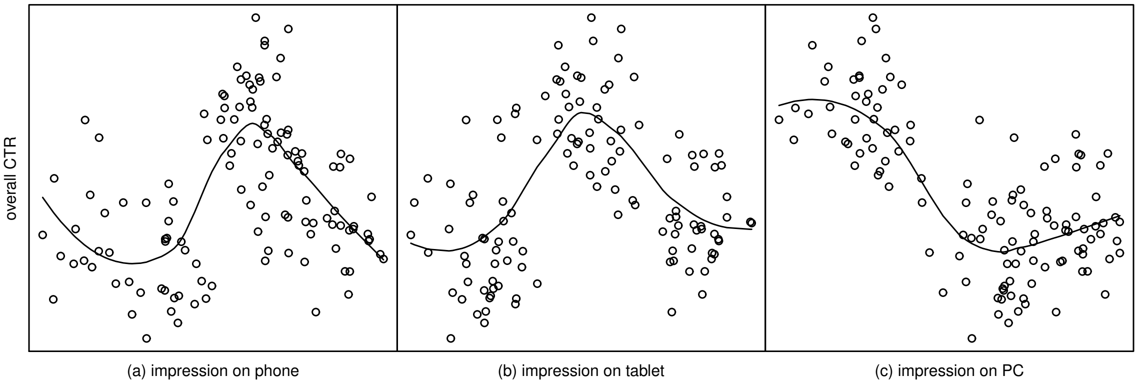

Figure 1. *The overall click-through rate v.s. the impression of a certain advertisement that is delivered on phone (left plot), tablet (middle plot), and PC (right plot), respectively. The black solid curves are the fitted locally weighted scatter-plot smoother (LOESS) curves.*

$\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_m}$, we consider a sparse tensor additive regression (STAR) model with

$$\mathcal{T}^*(\mathcal{X}_i) = \sum_{j_1=1}^{p_1} \cdots \sum_{j_m=1}^{p_m} f^*_{j_1 \ldots j_m}([\mathcal{X}_i]_{j_1 \ldots j_m}), \tag{1}$$

where $[\mathcal{X}_i]_{j_1 \ldots j_m}$ denotes the $(j_1, \ldots, j_m)$-th element of $\mathcal{X}_i$, and $f^*_{j_1 \ldots j_m}(\cdot)$ is a nonparametric additive component belonging to some smooth function class. Approximating the additive component $f^*_{j_1 \ldots j_m}(\cdot)$ using spline series expansions, $\mathcal{T}^*(\mathcal{X}_i)$ can be simplified to have a compact tensor representation of spline coefficients. To reduce the number of parameters and increase computational efficiency, we assume that the corresponding high-dimensional coefficient tensors have low-rank and group sparsity structures. Both low-rankness and sparsity are commonly used dimension reduction tools in recent tensor models (Li and Zhang, 2017; Sun et al., 2017; Sun and Li, 2017; Hao et al., 2018; Zhang, 2019; Zhang and Han, 2019). Besides effectively reducing computational cost, the group sparsity structure also significantly improves the model interpretability. For instance, in the online advertising example, when the daily overall CTR is regressed on the impressions of different advertisements on different devices, the group sparsity enables our STAR model to select effective advertisement and device combinations. Such a type of advertisement selection is important for managerial decision making and has been an active research area (Choi et al., 2010; Xu et al., 2016). To efficiently estimate the model, we formulate the parameter estimation as a non-convex optimization and propose a penalized alternating minimization algorithm. By fully exploiting the low-rankness and group sparsity structures as well as developing an efficient algorithm, our STAR model may run faster than the tensor linear regression in some experiments. For example, in the online advertising application, our STAR model can reduce the CTR prediction error by 50% while using 10% computational time of the linear or nonlinear tensor regression benchmark models. See Section 6 for more details.

Besides methodological contributions, we also obtain some strong theoretical results for our proposed method. In particular, we first establish a general theory for penalized alternating minimization in the context of tensor additive model. To the best of our knowledge, this is the first statistical-versus-optimization guarantee for the penalized alternating minimization. Previous work mostly focus on either the EM-type update (Wang et al., 2014; Balakrishnan et al., 2017; Hao et al., 2017), or the truncation-based update (Sun et al., 2017). Those techniques are not directly applicable to our scenario; see Section 4.1 for detailed explanations. Next, we derive a non-asymptotic error bound for the estimator from each iteration, which demonstrates the improvement of the estimation error in each update. Finally, we apply this general theory to our STAR estimator with B-spline basis and the group-lasso penalty, and show that the estimation error in the $(t+1)$-th iteration satisfies

$$\mathcal{E}^{(t+1)} \leq \underbrace{\rho^{t+1}\mathcal{E}^{(0)}}_{\text{optimization error}} + \underbrace{\frac{C_1}{1-\rho} n^{-\frac{2\kappa-1}{2\kappa+1}} \log(pd_n)}_{\text{statistical error}},$$

where $0 < \rho \leq 1/2$ is a contraction parameter, κ is the smoothness parameter of the function class, $p = \max\{p_1, \ldots, p_m\}$, and d_n is the number of spline series. The above error bound reveals an interesting interplay between the optimization error and the statistical error. The optimization error decays geometrically with the iteration number t, while the statistical error remains the same as t grows. When the tensor covariate is of order one (i.e., a vector covariate), our problem reduces to the vector nonparametric additive model. In that case, our statistical error matches with that from the vector nonparametric additive model in Huang et al. (2010).

1.1 Other related work

The problem we consider in our work is fundamentally different from those in tensor decomposition and tensor response regression. As a result, the technical tools involved and the theoretical results are quite different.

Tensor decomposition (Chi and Kolda, 2012; Anandkumar et al., 2014; Yuan and Zhang, 2016; Sun et al., 2017) is an unsupervised learning method that aims to find the best low-rank approximation of a single tensor. In comparison, our STAR model is a supervised learning method that seeks to capture the nonlinear relationship between the response and the tensor covariate. Although the low-rank structure of the tensor coefficient is also employed in our estimation, our objective and the technical tools involved are entirely different from the typical tensor decomposition problem. Additionally, one fundamental difference is that our model works with multiple tensor samples, while tensor decomposition works only with a single tensor. As a result, our error bound is a function of the sample size, which is different from that in tensor decomposition.

Another line of related work considers tensor response regression, where the response is a tensor and the covariates are scalars (Zhu et al., 2009; Li and Zhang, 2017; Sun and Li, 2017). These work also utilized the low-rank and/or sparse structures of the coefficient tensors for dimension reduction. However, tensors are treated as the *response* in tensor response regression, whereas they are treated as a *covariates* in our approach. These are two very different types of models, motivated by different applications. The tensor response regression aims to study the change of the tensor (e.g., the brain image) as the covariate (e.g., disease status) varies. However, the tensor regression model focuses on understanding the change of a scalar outcome (e.g., the overall CTR) with the tensor covariates. As a result, technical tools used for theoretical analysis are also largely different.

1.2 Notations and structure

Throughout this article, we denote scalars by lower case characters such as x, vectors by lower-case bold characters such as $\boldsymbol{x}$, matrices by upper-case bold characters such as $\boldsymbol{X}$ and tensors by upper-case script characters such as $\mathcal{X}$. Given a vector $\boldsymbol{x} \in \mathbb{R}^p$ and a set of indices $T \subset \{1, \ldots, p\}$, we define $\boldsymbol{x}_T$ such that $x_{T_j} = x_j$ if $j \in T$ and $x_{T_j} = 0$, otherwise. For a square matrix $\boldsymbol{A}$, we denote $\sigma_{\min}(\boldsymbol{A})$ and $\sigma_{\max}(\boldsymbol{A})$ as its minimum and maximum eigenvalues, respectively. For any function f on $[a, b]$, we define its $\ell_2(P)$ norm by $\|f(x)\|_2 = \sqrt{\int_a^b f^2(x) dP(x)}$. Suppose $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ are m-way tensors. We define tensor inner product $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{j_1, \ldots, j_m} \mathcal{X}_{j_1 \cdots j_m} \mathcal{Y}_{j_1 \cdots j_m}$. The tensor Frobenius norm is defined as $\|\mathcal{X}\|_F = \sqrt{\sum_{j_1=1}^{p_1} \cdots \sum_{j_m=1}^{p_m} \mathcal{X}^2_{j_1 \cdots j_m}}$. The notation $a \lesssim b$ implies $a \leq C_1 b$ for some constant $C_1 > 0$. For any two sequences $\{a_n\}_{n=1}^{\infty}, \{b_n\}_{n=1}^{\infty}$, we write $a_n = \mathcal{O}(b_n)$ if there exists some positive constant C_2 and sufficiently large n such that $a_n \leq C_2 b_n$. We also write $a_n \asymp b_n$ if there exist constants $C_3, C_4 > 0$ such that $C_3 a_n \leq b_n \leq C_4 a_n$ for all $n \geq 1$.

The rest of the article is organized as follows. Section 2 introduces our sparse tensor additive regression model. Section 3 develops an efficient penalized alternating minimization algorithm for model estimation. Section 4 investigates its theoretical properties, followed by simulation studies in Section 5 and a real online advertising application in Section 6. The appendix collects all technical proofs.

2. Sparse Tensor Additive Model

Given i.i.d. samples $\{y_i, \mathcal{X}_i\}_{i=1}^n$, our sparse tensor additive model assumes

$$y_i = \mathcal{T}^*(\mathcal{X}_i) + \epsilon_i = \sum_{j_1=1}^{p_1} \cdots \sum_{j_m=1}^{p_m} f^*_{j_1 \cdots j_m}([\mathcal{X}_i]_{j_1 \cdots j_m}) + \epsilon_i, \ i = 1, \ldots n, \tag{2}$$

where $f^*_{j_1 \cdots j_m}(\cdot)$ is the nonparametric additive function belonging to some smooth function class $\mathcal{H}$, and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. observation noises.

Our STAR model utilizes spline series expansion (Huang et al., 2010; Fan et al., 2011) to approximate each individual nonparametric additive component. Let $\mathcal{S}_n$ be the space of polynomial splines and $\{\psi_h(x)\}_{h=1}^{d_n}$ be a normalized basis for $\mathcal{S}_n$, where d_n is the number of spline series and $\sup_x |\psi_h(x)| \leq 1$. It is known that for any $f_n \in \mathcal{S}_n$, there always exists some coefficients $\{\beta_h^*\}_{h=1}^{d_n}$ such that $f_n(x) = \sum_{h=1}^{d_n} \beta_h^* \psi_h(x)$. In addition, under suitable smoothness assumptions (see Lemma 28), each nonparametric additive component $f^*_{j_1 \cdots j_m}(\cdot)$ can be well approximated by functions in $\mathcal{S}_n$. Applying the above approximation to each individual component, the regression function $\mathcal{T}^*(\mathcal{X}_i)$ in (2) can be approximated by

$$\mathcal{T}^*(\mathcal{X}_i) \approx \sum_{j_1=1}^{p_1} \cdots \sum_{j_m=1}^{p_m} \sum_{h=1}^{d_n} \beta^*_{j_1 \cdots j_m h} \psi_{j_1 \cdots j_m h}([\mathcal{X}_i]_{j_1 \cdots j_m}). \tag{3}$$

The expression in (3) has a compact tensor representation. Define $\mathcal{F}_h(\mathcal{X}) \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ such that $[\mathcal{F}_h(\mathcal{X})]_{j_1 \cdots j_m} = \psi_{j_1 \cdots j_m h}([\mathcal{X}]_{j_1 \cdots j_m})$, and $\mathcal{B}_h^* \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ such that $[\mathcal{B}_h^*]_{j_1 \cdots j_m} = \beta^*_{j_1 \cdots j_m h}$

for $h \in [d_n]$, where $[k]$ denotes $\{1, \dots, k\}$ for an integer $k \geq 1$. Consequently, we can write

$$\sum_{j_1=1}^{p_1} \cdots \sum_{j_m=1}^{p_m} \sum_{h=1}^{d_n} \beta^*_{j_1 \dots j_m h} \psi_{j_1 \dots j_m h}([\mathcal{X}_i]_{j_1 \dots j_m}) = \sum_{h=1}^{d_n} \Big\langle \mathcal{B}^*_h, \mathcal{F}_h(\mathcal{X}_i) \Big\rangle. \tag{4}$$

Therefore, the parameter estimation of the nonparametric additive model (2) reduces to the estimation of unknown tensor coefficients $\mathcal{B}^*_1, \dots, \mathcal{B}^*_{d_n}$. The coefficients $\mathcal{B}^*_1, \dots, \mathcal{B}^*_{d_n}$ include a total number of $\mathcal{O}(d_n \Pi_{j=1}^m p_j)$ free parameters, which could be much larger than the sample size n. In such ultrahigh-dimensional scenario, it is important to employ dimension reduction tools. A common tensor dimension reduction tool is the low-rank assumption (Chi and Kolda, 2012; Anandkumar et al., 2014; Yuan and Zhang, 2016; Sun et al., 2017). Similarly, we assume each coefficient tensor $\mathcal{B}^*_1, \dots, \mathcal{B}^*_{d_n}$ satisfies the CP low-rank decomposition (Kolda and Bader, 2009):

$$\mathcal{B}^*_h = \sum_{r=1}^{R} \boldsymbol{\beta}^*_{1hr} \circ \cdots \circ \boldsymbol{\beta}^*_{mhr}, \ h = 1, \dots, d_n, \tag{5}$$

where $\circ$ is the vector outer product, $\boldsymbol{\beta}^*_{1hr} \in \mathbb{R}^{p_1}, \dots, \boldsymbol{\beta}^*_{mhr} \in \mathbb{R}^{p_m}$, and $R \ll \min\{p_1, \dots, p_m\}$ is the CP-rank. This formulation reduces the effective number of the parameters from $\mathcal{O}(d_n \Pi_{j=1}^m p_j)$ to $\mathcal{O}(d_n R \sum_{j=1}^m p_j)$, and hence greatly improves computational efficiency. Under this formulation, our model can be written as

$$\mathcal{T}^*(\mathcal{X}_i) \approx \sum_{h=1}^{d_n} \Big\langle \mathcal{F}_h(\mathcal{X}_i), \sum_{r=1}^{R} \boldsymbol{\beta}^*_{1hr} \circ \cdots \circ \boldsymbol{\beta}^*_{mhr} \Big\rangle. \tag{6}$$

Remark 1 *Our model in* (6) *can be viewed as a generalization of several existing work. When* $\psi_{j_1 \dots j_m h}(\cdot)$ *in* (4) *is an identity basis function* ($\psi_{j_1 \dots j_m h}([\mathcal{X}]_{j_1 \dots j_m}) = [\mathcal{X}]_{j_1 \dots j_m}$) *with only one basis* ($d_n = 1$), (6) *reduces to the bilinear form (Li et al., 2010; Hung and Wang, 2012) for a matrix covariate* ($m = 2$)*, and the multilinear form for linear tensor regression (Zhou et al., 2013; Hoff, 2015; Yu and Liu, 2016; Rabusseau and Kadri, 2016; Sun and Li, 2017; Guhaniyogi et al., 2017; Raskutti et al., 2019) for a tensor covariate* ($m \geq 3$)*.*

In addition to the CP low-rank structure on the tensor coefficients, we further impose a group-type sparsity constraint on the components $\boldsymbol{\beta}^*_{khr}$. This group sparsity structure not only further reduces the effective parameter size, but also improves the model interpretability, as it enables the variable selection of components in the tensor covariate. Recall that in (6) we have $\boldsymbol{\beta}^*_{khr} = (\beta^*_{khr1}, \dots, \beta^*_{khrp_k})^\top$ for $k \in [m], h \in [d_n], r \in [R]$. We define our group sparsity constraint as

$$\Bigg| \underbrace{\Big\{ j \in [p_k] \Big| \sum_{h=1}^{d_n} \sum_{r=1}^{R} \beta^{*2}_{khrj} \neq 0 \Big\}}_{\mathcal{S}_k} \Bigg| = s_k \ll p_k, \ \text{for } k \in [m]. \tag{7}$$

where $|\mathcal{S}_k|$ refers to the cardinality of the set $\mathcal{S}_k$. Figure 2 provides an illustration of the low-rank (5) and group-sparse (7) coefficients when the order of the tensor is $m = 3$. When $m = 1$, our model with the group sparsity constraint reduces to the vector sparse additive model (Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010).

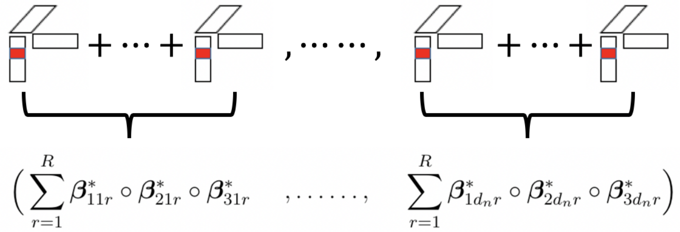
$\Big(\sum_{r=1}^{R}\boldsymbol{\beta}_{11r}^{*}\circ\boldsymbol{\beta}_{21r}^{*}\circ\boldsymbol{\beta}_{31r}^{*}$
$,\ldots\ldots,$
$\sum_{r=1}^{R}\boldsymbol{\beta}_{1d_nr}^{*}\circ\boldsymbol{\beta}_{2d_nr}^{*}\circ\boldsymbol{\beta}_{3d_nr}^{*}\Big)$

Figure 2. *An illustration of the low-rank and group-sparsity structures in a collection of three-way tensor coefficients* $(\mathcal{B}_1^*,\ldots,\mathcal{B}_{d_n}^*)$*. If one or more of the coefficients at the colored locations are non-zero, the cardinality of* $\mathcal{S}_1$ *increases by one.*

Remark 2 *Consider the tensor order as 2. The model in* (6) *reduces to*

$$\mathcal{T}^*(\mathcal{X}_i) \approx \sum_{h=1}^{d_n} \Big\langle \mathcal{F}_h(\mathcal{X}_i), \sum_{r=1}^{R} \boldsymbol{\beta}_{1hr}^{*}\boldsymbol{\beta}_{2hr}^{*\top} \Big\rangle \in \mathbb{R}^{p_1\times p_2}.$$

For instance, suppose $\sum_{h=1}^{d_n}\sum_{r=1}^{R}\beta_{1hr1}^{*2}=0$*. This implies the first row of RHS is all 0 and thus encourages variable selection in* $\mathcal{T}^*(\mathcal{X}_i)$ *correspondingly.*

3. Estimation

In this section, we describe our approach to estimate the parameters in our STAR model via a penalized empirical risk minimization which simultaneously satisfies the low-rankness and encourages the sparsity of decomposed components. In particular, we consider

$$\min_{\boldsymbol{\beta}_{1hr},\ldots,\boldsymbol{\beta}_{mhr}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(y_i-\sum_{h=1}^{d_n}\Big\langle\sum_{r=1}^{R}\boldsymbol{\beta}_{1hr}\circ\cdots\circ\boldsymbol{\beta}_{mhr},\mathcal{F}_h(\mathcal{X}_i)\Big\rangle\Big)^2}_{\mathcal{L}(\boldsymbol{\beta}_{1hr},\ldots,\boldsymbol{\beta}_{mhr})}+\mathcal{P}(\boldsymbol{\beta}_{1hr},\ldots,\boldsymbol{\beta}_{mhr}), \qquad (8)$$

where $\mathcal{L}(\boldsymbol{\beta}_{1hr},\ldots,\boldsymbol{\beta}_{mhr})$ is the empirical risk function, in which the low-rankness is guaranteed due to the CP decomposition, and $\mathcal{P}(\cdot)$ is a penalty term that encourages sparsity. To enforce the sparsity as defined in (7), we consider the group lasso penalty (Yuan and Lin, 2006), i.e.,

$$\mathcal{P}(\boldsymbol{\beta}_{1hr},\ldots,\boldsymbol{\beta}_{mhr}) = \sum_{k=1}^{m}\Big(\lambda_{kn}\sum_{j=1}^{p_k}\sqrt{\sum_{h=1}^{d_n}\sum_{r=1}^{R}\beta_{khrj}^2}\Big), \qquad (9)$$

where $\{\lambda_{kn}\}_{k=1}^{m}$ are tuning parameters. It is worth mentioning that our algorithm and theoretical analysis can accommodate a general class of decomposable penalties (see Condition 12 for details), which includes lasso, ridge, fused lasso, and group lasso as special cases.

For a general tensor covariate ($m>1$), the optimization problem in (8) is a non-convex optimization. This is fundamentally different from the vector sparse additive model (Ravikumar et al., 2009; Huang et al., 2010) whose optimization is convex. The non-convexity in (8) brings significant challenges in both model estimation and theoretical development. The key idea of our estimation procedure is to explore the bi-convex structure of the empirical risk function $\mathcal{L}(\boldsymbol{\beta}_{1hr}, \ldots, \boldsymbol{\beta}_{mhr})$ since it is convex in one argument while fixing all the other parameters. This motivates us to rewrite the empirical risk function into a bi-convex representation, which in turn facilitates the introduction of an efficient alternating minimization algorithm.

Denote $\vartheta_{krj} = (\beta_{k1rj}, \beta_{k2rj}, \ldots, \beta_{kd_nrj})^\top \in \mathbb{R}^{d_n \times 1}, \vartheta_{kj} = (\vartheta_{k1j}^\top, \ldots, \vartheta_{kRj}^\top)^\top \in \mathbb{R}^{Rd_n \times 1}$ for $k \in [m]$, $j \in [p_k]$, and $\boldsymbol{b}_k = (\vartheta_{k1}^\top, \ldots, \vartheta_{kp_k}^\top)^\top$. We also define the operator $\prod_{k\in[m]}^{\circ} \boldsymbol{a}_k = \boldsymbol{a}_1 \circ \cdots \circ \boldsymbol{a}_m$. Remind that $\mathcal{F}_h(\mathcal{X}) \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ with $[\mathcal{F}_h(\mathcal{X})]_{j_1 \cdots j_m} = \psi_{j_1 \cdots j_m h}([\mathcal{X}]_{j_1 \cdots j_m})$, see (4). We use $[\mathcal{F}_h^k(\mathcal{X}_i)]_j$ to refer to the $m-1$ way tensor when we fix the index along the k-th way of $\mathcal{F}_h(\mathcal{X}_i)$ as j, e.g., $[\mathcal{F}_h^1(\mathcal{X}_i)]_j \in \mathbb{R}^{p_2 \times \cdots \times p_m}$. Define

$$F_{irj}^k = \Big(\big\langle \prod_{u\in[m]\backslash k}^{\circ} \boldsymbol{\beta}_{u1r}, [\mathcal{F}_1^k(\mathcal{X}_i)]_j \big\rangle, \ldots, \big\langle \prod_{u\in[m]\backslash k}^{\circ} \boldsymbol{\beta}_{ud_nr}, [\mathcal{F}_{d_n}^k(\mathcal{X}_i)]_j \big\rangle\Big)^\top,$$

and denote $F_{ij}^k = (F_{i1j}^{k\top}, \ldots, F_{iRj}^{k\top})^\top \in \mathbb{R}^{Rd_n \times 1}$. In addition, we denote $\boldsymbol{F}_j^k = (F_{1j}^k, \ldots, F_{nj}^k)^\top$, $\boldsymbol{F}^k = (\boldsymbol{F}_1^k, \ldots, \boldsymbol{F}_{p_k}^k)$, and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Thus, when other parameters are fixed, minimizing the empirical risk function (8) with respect to $\boldsymbol{b}_k$ is equivalent to minimizing

$$\mathcal{L}\big(\boldsymbol{b}_1, \cdots, \boldsymbol{b}_m\big) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{F}^k \boldsymbol{b}_k\|_2^2. \tag{10}$$

Note that the expression of (10) holds for any $k \in [m]$ with proper definitions on $\boldsymbol{F}^k$ and $\boldsymbol{b}_k$.

Remark 3 *Intuitively, ϑ_{kj} summarizes all the colored coefficients in Figure 2 and $\boldsymbol{b}_k$ summarizes all the coefficients along k-th mode. By this definition, we can more clearly describe the effect of group sparsity: when ϑ_{kj} is a zero vector, the j-th variable in k-th mode is irrelevant.*

Based on this reformulation, we are ready to introduce the alternating minimization algorithm that solves (8) by alternatively updating $\boldsymbol{b}_1, \cdots, \boldsymbol{b}_m$. A desirable property of our algorithm is that updating $\boldsymbol{b}_k$ given others can be solved efficiently via the group-wise coordinate descent based on the back-fitting algorithm (Ravikumar et al., 2009). The detailed algorithm is summarized in Algorithm 1. With a little abuse of notations, we redefine the penalty term $\mathcal{P}(\boldsymbol{b}_k^{(t)}) = \sum_{j=1}^{p_k} \sqrt{\sum_{h=1}^{d_n} \sum_{r=1}^{R} (\beta_{khrj}^{(t)})^2}$.

In our implementation, we use ridge regression to initialize Algorithm 1, and set tuning parameters $\lambda = \lambda_{kn}$ for $k = 1, \ldots, m$ for simplicity. When solving Problem (11), we use the *warm start* and *active set* to accelerate the algorithm. The basic idea of the two tricks is illustrated as follows: for each t, we use the solution $\boldsymbol{b}_k^{(t-1)}$ as the initial value to update $\boldsymbol{b}_k^{(t)}$, i.e., the warm start; when computing $\boldsymbol{b}_k^{(t)}$, we may only consider an active set of k such that $\boldsymbol{b}_k^{(t-1)}$ is nonzero. Those two tricks have shown great successes in the implementations of the coordinate-descent-type algorithms such as the R package `glmnet` (Friedman et al., 2010). The overall computational complexity of Algorithm 1 is $O(Tmkn(Rp)^m)$, where T is the number of iterations for alternating minimization, m is the order the tensor, k is the number of iteration for group-wise coordinate descent to solve Problem (11), n is the sample size, R is the tensor rank and p is the maximum dimension of each tensor mode.

Algorithm 1 Penalized Alternating Minimization for Solving (8)

```
Input: $\{y_i\}_{i=1}^n$, $\{\mathcal{X}_i\}_{i=1}^n$, initialization $\{\boldsymbol{b}_1^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\}$, the set of penalization parameters $\{\lambda_{1n}, \ldots, \lambda_{mn}\}$, rank $R$, iteration $t = 0$, stopping error $\epsilon = 10^{-5}$.
Repeat $t = t + 1$ and run penalized alternating minimization.
    For $k = 1$ to $m$
        $\boldsymbol{b}_k^{(t+1)} = \underset{\boldsymbol{b}_k}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{b}_1^{(t)}, \ldots, \boldsymbol{b}_m^{(t)}) + \lambda_{kn}\mathcal{P}(\boldsymbol{b}_k^{(t)}),$   (11)
        where $\mathcal{L}$ is defined in (10).
    End for.
Until $\max_k \|\boldsymbol{b}_k^{(t+1)} - \boldsymbol{b}_k^{(t)}\|_2 \leq \epsilon$ , and let $t = T^*$.
Output: the estimate of each component, $\{\boldsymbol{b}_1^{(T^*)}, \ldots, \boldsymbol{b}_m^{(T^*)}\}$.
```

4. Theory

In this section, we first establish a general theory for the penalized alternating minimization in the context of the tensor additive model. Several sufficient conditions are proposed to guarantee the optimization error and statistical error. Then, we apply our theory to the STAR estimator with B-spline basis functions and the group-lasso penalty. To ease the presentation, we consider a three-way tensor covariate (i.e, $m = 3$) in our theoretical development, while its generalization to an m-way tensor is straightforward.

4.1 A general contract property

To bound the optimization error and statistical error of the proposed estimator, we introduce three sufficient conditions: a Lipschitz-gradient condition, a sparse strongly convex condition, and a generic statistical error condition. For the sake of brevity, we only present conditions for the update of $\boldsymbol{b}_1$ in the main paper, and defer similar conditions for $\boldsymbol{b}_2, \boldsymbol{b}_3$ to Section 2 in the appendix.

For each vector $\boldsymbol{x} \in \mathbb{R}^{pd_nR \times 1}$, we divide it into p equal-length segments as in Figure 3. A segment is colored if it contains at least one non-zero element, and a segment is uncolored if all of its elements are zero. We let $w(\boldsymbol{x})$ be the indices of colored segments in $\boldsymbol{x}$ and E_s be the set of all (pd_nR)-dimensional vectors with less than $C_0 s$ colored segments, for some constant C_0. Mathematically, for a vector $\boldsymbol{x} \in \mathbb{R}^{pd_nR \times 1}$, denote $w(\boldsymbol{x}) := \{j \in [p] | \sum_{h=1}^{d_nR} x_{(j-1)d_nR+h}^2 \neq 0\}$ and $E_s := \{\boldsymbol{x} \in \mathbb{R}^{pd_nR \times 1} || w(\boldsymbol{x})| \leq C_0 s\}$.

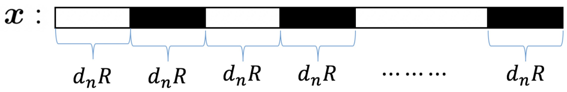

Figure 3. *An illustration of the group sparse vector. A segment is colored if it contains at least one non-zero element, and a segment is uncolored if all of its elements are zero.*

Define a sparse ball $\mathcal{B}_{\alpha,s}(\boldsymbol{b}^*) := \{\boldsymbol{b} \in \mathbb{R}^{pd_nR} : \|\boldsymbol{b} - \boldsymbol{b}^*\|_2 \leq \alpha, \boldsymbol{b} \in E_s\}$ for a given constant radius α. Moreover, the noisy gradient function and noiseless gradient function of empirical risk function $\mathcal{L}$ defined in (8) of order-3 with respect to $\boldsymbol{b}_1$ can be written as

$$\begin{aligned} \nabla_1 \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) &= \frac{2}{n} \boldsymbol{F}^{1\top} \Big(\boldsymbol{F}^1 \boldsymbol{b}_1 - \boldsymbol{y} \Big) \\ \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) &= \frac{2}{n} \boldsymbol{F}^{1\top} \Big(\boldsymbol{F}^1 \boldsymbol{b}_1 - \boldsymbol{F}^{1*} \boldsymbol{b}_1^* \Big), \end{aligned} \tag{12}$$

where $F_{irj}^{1*} = \big(\langle \boldsymbol{\beta}_{21r}^* \circ \boldsymbol{\beta}_{31r}^*, [\mathcal{F}_1^1(\mathcal{X}_i)]_j \rangle, \ldots, \langle \boldsymbol{\beta}_{2d_nr}^* \circ \boldsymbol{\beta}_{3d_nr}^*, [\mathcal{F}_{d_n}^1(\mathcal{X}_i)]_j \rangle \big)^\top$.

Condition 4 (Lipschitz-Gradient) *For $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2, s_2}(\boldsymbol{b}_2^*), \boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3, s_3}(\boldsymbol{b}_3^*)$, the noiseless gradient function $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \cdot, \boldsymbol{b}_3^*)$ satisfies μ_{2n}-Lipschitz-gradient condition, and $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \cdot)$ satisfies μ_{3n}-Lipschitz-gradient condition with high probability. That is,*

$$\begin{aligned} \big\langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\rangle &\le \mu_{2n} \big\| \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\|_2 \big\| \boldsymbol{b}_2 - \boldsymbol{b}_2^* \big\|_2 \\ \big\langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*), \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\rangle &\le \mu_{3n} \big\| \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\|_2 \big\| \boldsymbol{b}_3 - \boldsymbol{b}_3^* \big\|_2, \end{aligned}$$

with probability at least $1 - \delta_1$ for any $0 < \delta_1 < 1$. Here, μ_{2n}, μ_{3n} may depend on δ_1.

Remark 5 *Condition 4 defines a variant of Lipschitz continuity for $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \cdot, \boldsymbol{b}_3^*)$ and $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \cdot)$. Note that the gradient is always taken with respect to the first argument of $\mathcal{L}(\cdot, \cdot, \cdot)$ and the Lipschitz continuity is with respect to the second or the third argument. Analogous Lipschitz-gradient conditions were also considered in Balakrishnan et al. (2017); Hao et al. (2017) for the* population-level Q-function *in the EM-type update.*

Next condition characterizes the curvature of noisy gradient function in a sparse ball. It states that when the second and the third argument are fixed, $\mathcal{L}(\cdot, \cdot, \cdot)$ is strongly convex with parameter γ_{1n} with high probability. As shown later in Section 4.2, this condition holds for a broad family of basis functions.

Condition 6 (Sparse-Strong-Convexity) *For any $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2, s_2}(\boldsymbol{b}_2^*), \boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3, s_3}(\boldsymbol{b}_3^*)$, the loss function $\mathcal{L}(\cdot, \cdot, \cdot)$ is sparse strongly convex in its first argument, namely*

$$\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_1 \mathcal{L}\big(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^* - \boldsymbol{b}_1 \rangle \ge \frac{\gamma_{1n}}{2} \| \boldsymbol{b}_1 - \boldsymbol{b}_1^* \|_2^2,$$

with probability at least $1 - \delta_2$ for any $0 < \delta_2 < 1$. Here, $\gamma_{1n} > 0$ is the strongly convex parameter and may depend on δ_2.

Next we present the definition for dual norm, which is a key measure for statistical error condition. More details on the dual norm are referred to Negahban et al. (2012).

Definition 7 (Dual norm) *For a given inner product $\langle \cdot, \cdot \rangle$, the dual norm of $\mathcal{P}$ is given by*

$$\mathcal{P}^*(\boldsymbol{v}) := \sup_{\boldsymbol{u} \in \mathbb{R}^p \setminus \{0\}} \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\mathcal{P}(\boldsymbol{u})}.$$

As a concrete example, the dual of ℓ_1-norm is ℓ_∞-norm while the dual of ℓ_2-norm is itself. Suppose $\boldsymbol{v}$ is a p-dimensional vector and the index set $\{1, 2, \ldots, p\}$ is partitioned into $N_{\mathcal{G}}$ disjoint groups, namely $\mathcal{G} = \{G_1, \ldots, G_{N_{\mathcal{G}}}\}$. The group norm for $\boldsymbol{v}$ is defined as $\mathcal{P}(\boldsymbol{v}) = \sum_{t=1}^{N_{\mathcal{G}}} \|\boldsymbol{v}_{G_t}\|_2$. According to Definition (7), the dual of $\mathcal{P}(\boldsymbol{v})$ is defined as $\mathcal{P}^*(\boldsymbol{v}) = \max_t \|\boldsymbol{v}_{G_t}\|_2$. For simplicity, we write $\|\cdot\|_{\mathcal{P}^*} = \mathcal{P}^*(\cdot)$.

The generic statistical error (SE) condition guarantees that the distance between noisy gradient and noiseless gradient under $\mathcal{P}^*$-norm is bounded.

Condition 8 (Statistical-Error) *For any $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$, we have with probability at least $1-\delta_3$,*

$$\left\|\nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\right\|_{\mathcal{P}^*} \le \varepsilon_1.$$

Remark 9 *Here, ε_1 is only a generic quantity and its explicit form will be derived for a specific loss function in Section 4.2.*

Next we introduce two conditions for the penalization parameter (Condition 10) and penalty (Condition 12). To illustrate Condition 10, we first introduce an quantity called *support space compatibility constant* to measure the intrinsic dimensionality of $\mathcal{S}_1$ defined in (7) with respect to penalty $\mathcal{P}$. Specifically, it is defined as

$$\Phi(\mathcal{S}_1) := \sup_{\boldsymbol{b} \in \mathcal{S}_1 \backslash \{0\}} \frac{\mathcal{P}(\boldsymbol{b})}{\|\boldsymbol{b}\|_2}, \tag{13}$$

which is a variant of subspace compatibility constant originally proposed by Negahban et al. (2012) and Wainwright (2014). If $\mathcal{P}(\boldsymbol{b})$ is chosen as a group lasso penalty, we have $\Phi(\mathcal{S}') = \sqrt{|\mathcal{S}'|}$, where $\mathcal{S}'$ is the index set of active groups. Similar definitions of $\Phi(\mathcal{S}_2), \Phi(\mathcal{S}_3)$ can be made accordingly.

Condition 10 (Penalization Parameter) *We consider an iterative turning procedure where tuning parameters in (11) are allowed to change with iteration. In particular, we assume tuning parameters $\{\lambda_{1n}^{(t)}, \lambda_{2n}^{(t)}, \lambda_{3n}^{(t)}\}$ satisfy*

$$\begin{aligned}
\lambda_{1n}^{(t)} &= 4\varepsilon_1 + (\mu_{2n}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 + \mu_{3n}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2)/\Phi(\mathcal{S}_1)\\
\lambda_{2n}^{(t)} &= 4\varepsilon_2 + (\mu_{1n}'\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2 + \mu_{3n}^{'}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2)/\Phi(\mathcal{S}_2)\\
\lambda_{3n}^{(t)} &= 4\varepsilon_3 + (\mu_{1n}^{''}\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2 + \mu_{2n}^{''}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2)/\Phi(\mathcal{S}_3),
\end{aligned}$$

where $\{\mu_{2n}, \mu_{3n}\}, \{\mu_{1n}', \mu_{3n}'\}, \{\mu_{1n}^{''}, \mu_{2n}^{''}\}$ are Lipschitz-gradient parameter which are defined in Condition 4 and Conditions 29-30.

Remark 11 *Condition 10 considers an iterative sequence of regularization parameters. Given reasonable initializations for $\boldsymbol{b}_1^{(t)}, \boldsymbol{b}_2^{(t)}, \boldsymbol{b}_3^{(t)}$, their estimation errors gradually decay when the iteration t increases, which implies that $\lambda_{kn}^{(t)}$ is a decreasing sequence. After sufficiently many iterations, the rate of the $\lambda_{kn}^{(t)}$ will be bounded by the statistical error ε_k, for $k=1,2,3$. This agrees with the theory of high-dimensional regularized M-estimator in that suitable tuning parameter should be proportional to the target estimation error (Wainwright, 2014). Such iterative turning procedure plays a critical role in controlling statistical and optimization error, and has been commonly used in other high-dimensional non-convex optimization problems (Wang et al., 2014; Yi and Caramanis, 2015).*

Finally, we present a general condition on the penalty term.

Condition 12 (Decomposable Penalty) *Given a space $\mathcal{S}$, a norm-based penalty $\mathcal{P}$ is assumed to be decomposable with respect to $\mathcal{S}$ such that it satisfies $\mathcal{P}(\boldsymbol{u}+\boldsymbol{v}) = \mathcal{P}(\boldsymbol{u}) + \mathcal{P}(\boldsymbol{v})$ for any $\boldsymbol{u} \in \mathcal{S}$ and $\boldsymbol{v} \in \mathcal{S}^{\perp}$, where $\mathcal{S}^{\perp}$ is the complement pf $\mathcal{S}$.*

As shown in Negahban and Wainwright (2011), a broad class of penalties satisfies the decomposable property, such as lasso, ridge, fused lasso, group lasso penalties. Next theorem quantifies the error of one-step update for the estimator coming Algorithm 1.

Theorem 13 (Contraction Property) *Suppose Conditions 4,6,8,10, 29-34 hold. Assume the update at t-th iteration of Algorithm 1, $\boldsymbol{b}_1^{(t)}, \boldsymbol{b}_2^{(t)}, \boldsymbol{b}_3^{(t)}$ fall into sparse balls $\mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*), \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*), \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$ respectively, where $\alpha_1, \alpha_2, \alpha_3$ are some constants. Define $\mathcal{E}^{(t)} = \|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2^2 + \|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2^2 + \|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2^2$. There exists absolute constant $C_0 > 1$ such that, the estimation error of the update at the $t+1$-th iteration satisfies,*

$$\mathcal{E}^{(t+1)} \le \rho \mathcal{E}^{(t)} + C_0 \Big(\frac{\varepsilon_1^2 \Phi(\mathcal{S}_1)^2}{\gamma_{1n}^2} + \frac{\varepsilon_2^2 \Phi(\mathcal{S}_2)^2}{\gamma_{2n}^2} + \frac{\varepsilon_3^2 \Phi(\mathcal{S}_3)^2}{\gamma_{3n}^2} \Big), \tag{14}$$

with probability at least $1 - 3(\delta_1 + \delta_2 + \delta_3)$. Here, ρ is the contraction parameter defined as

$$\rho = C_1 \max\{\mu_{1n}^{'2}, \mu_{1n}^{''2}, \mu_{2n}^2, \mu_{2n}^{''2}, \mu_{3n}^2, \mu_{3n}^{'2}\} / \min\{\gamma_{1n}^2, \gamma_{2n}^2, \gamma_{3n}^2\},$$

where C_1 is some constant.

Theorem 13 demonstrates the mechanism of how the estimation error improves in the one-step update. When the the contraction parameter ρ is strictly less than 1 (we will prove that it holds for certain class of basis functions and penalties in next section), the first term of RHS in (14) will gradually go towards zero and the second term will be stable. The contraction parameter ρ is roughly the ratio of Lipschitz-gradient parameter and the strongly convex parameter. Similar formulas of contraction parameter frequently appears in the literature of statistical guarantees for low/high-dimensional non-convex optimization (Balakrishnan et al., 2017; Hao et al., 2017).

Remark 14 *Yi and Caramanis (2015) provided similar optimization and statistical guarantee for regularized EM algorithm based on mixture model. However, the source of non-convexity in the mixture model comes from the latent variable while ours comes from the bi-convex structure in the low-rank model. Thus their analysis is not directly applicable to our case due to different verification of sparse-strong-convexity condition.*

4.2 Application to STAR estimator

In this section, we apply the general contract property in Theorem 13 to STAR estimator with B-spline basis functions. The formal definition of B-spline basis function is defined in Section 1 of the appendix. To ensure Conditions 4-6 and 8 are satisfied, in our STAR estimator we require conditions on the nonparametric component, the distribution of tensor covariate and the noise distribution.

Condition 15 (Function Class) *Each nonparametric component in* (1) *is assumed to belong to the function class $\mathcal{H}$ defined as follows,*

$$\mathcal{H} = \Big\{ g(\cdot) : |g^{(r)}(s) - g^{(r)}(t)| \leq C|s-t|^{\alpha}, \text{ for } s,t \in [a,b] \Big\}, \tag{15}$$

where r is the order of the derivative. Let $\kappa = r + \alpha > 0.5$ be the smoothness parameter of function class $\mathcal{H}$. For $j \in [p_1], k \in [p_2], l \in [p_3]$, there is a constant $c_f > 0$ such that $\min_{jkl} \|f^*_{jkl}(x)\|_2 \geq c_f$ and $\mathbb{E}(f^*_{jkl}([\mathcal{X}]_{jkl})) = 0$. Each component of the covariate tensor $\mathcal{X}$ has an absolutely continuous distribution and its density function is bounded away from zero and infinitely on C.

Condition 15 is classical for nonparametric additive model (Stone, 1985; Huang et al., 2010; Fan et al., 2011). Such condition is required to optimally estimate each individual additive component in ℓ_2-norm.

Condition 16 (Sub-Gaussian Noise) *The noise $\{\epsilon_i\}_{i=1}^n$ are i.i.d. randomly generated with mean 0 and bounded variance σ^2. Moreover, (ϵ_i/σ) is sub-Gaussian distributed, i.e., there exists constant $C_\epsilon > 0$ such that $\|(\epsilon_i/\sigma)\|_{\phi_2} := \sup_{p \geq 1} p^{-1/2}(\mathbb{E}|\epsilon_i/\sigma|^p)^{1/p} \leq C_\epsilon$, and independent of tensor covariates $\{\mathcal{X}_i\}_{i=1}^n$.*

Condition 17 (Parameter Space) *We assume the absolute value of maximum entry of $(\boldsymbol{b}_1^{*\top}, \boldsymbol{b}_2^{*\top}, \boldsymbol{b}_2^{*\top})$ is upper bounded by some positive constant c^*, and the absolute value of minimum non-zero entry of $(\boldsymbol{b}_1^{*\top}, \boldsymbol{b}_2^{*\top}, \boldsymbol{b}_2^{*\top})$ is lower bounded by some positive constant c_*. Here, c^*, c_* not depending on n, p. Moreover, we assume the CP-rank R and sparsity parameters s_1, s_2, s_3 are bounded by some constants.*

Remark 18 *The condition of bounded elements of tensor coefficient widely appears in tensor literature (Anandkumar et al., 2014; Sun et al., 2017). Here the bounded tensor rank condition is imposed purely for simplifying the proofs and this condition is possible to relax to allow slowly increased tensor rank (Sun and Li, 2019). The fixed sparsity assumption is also required in the vector nonparametric additive regression (Huang et al., 2010). To relax it, Meier et al. (2009) considered a diverging sparsity scenario but required a compatibility condition which was hard to verify in practice. Thus, in this paper we consider a fixed sparsity case and leave the extension of diverging sparsity as future work.*

Since the penalized empirical risk minimization (8) is a highly non-convex optimization, we require some conditions on the initial update in Algorithm 1.

Condition 19 (Initialization) *The initialization of $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ is assumed to fall into a sparse constant ball centered at $\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*$, saying $\boldsymbol{b}_1^{(0)} \in \mathcal{B}_{\alpha_1, s_1}(\boldsymbol{b}_1^*), \boldsymbol{b}_2^{(0)} \in \mathcal{B}_{\alpha_2, s_1}(\boldsymbol{b}_2^*), \boldsymbol{b}_3^{(0)} \in \mathcal{B}_{\alpha_3, s_1}(\boldsymbol{b}_3^*)$, where $\alpha_1, \alpha_2, \alpha_3$ are some constants that are not diverging with n or p.*

Remark 20 *Similar initialization conditions have been widely used in tensor decomposition (Sun et al., 2017; Sun and Li, 2019), tensor regression (Suzuki et al., 2016; Sun and Li, 2017), and other non-convex optimization (Wang et al., 2014; Yi and Caramanis, 2015). Once the initial values fall into the sparse ball, the contract property and group lasso ensure* *that the successive updates also fall into a sparse ball. Another line of work considers to design spectral methods to initialize certain simple non-convex optimization, such as matrix completion (Ma et al., 2017) and tensor sketching (Hao et al., 2018). The success of spectral methods heavily relies on a simple non-convex geometry and explicit form of high-order moment calculation, which is easy to achieve in previous work (Ma et al., 2017; Hao et al., 2018) by assuming a Gaussian error assumption. However, the design of spectral method in our tensor additive regression is substantially harder since the high-order moment calculation has no explicit form in our context. We leave the design of spectral-based initialization as future work.*

Finally, we state the main theory on the estimation error of our STAR estimator with B-spline basis functions and a group lasso penalty.

Theorem 21 *Suppose Conditions 10, 15-17, 19 hold and consider the class of normalized B-spline basis functions defined in* (A1) *and group-lasso penalty defined in* (9)*. If one chooses the number of spline series* $d_n \asymp n^{\frac{1}{2\kappa+1}}$ *and the tuning parameter* $\lambda_{1n}^{(t)}, \lambda_{2n}^{(t)}, \lambda_{3n}^{(t)}$ *as defined in Condition 10 with generic parameters specified in Lemmas 26-27, with probability at least* $1 - C_0(t+1)(Rsn^{-\frac{2\kappa}{2\kappa+1}} + 1/p)$*, we have*

$$\mathcal{E}^{(t+1)} \leq \underbrace{\rho^{t+1}\mathcal{E}^{(0)}}_{\textit{optimization error}} + \underbrace{\frac{C_1 R^2}{1-\rho} n^{-\frac{2\kappa-1}{2\kappa+1}} \log(pd_n)}_{\textit{statistical error}}, \tag{16}$$

where $0 < \rho \leq 1/2$ *is a contraction parameter, and* κ *is the smoothness parameter of function class* $\mathcal{H}$ *in* (15)*. Note that* C_1 *may involve a smaller order term that is negligible asymptotically. Consequently, when the total number of iterations is no smaller than*

$$T^* = \log\left(\frac{1-\rho}{C_1\mathcal{E}^{(0)}} \frac{n^{\frac{2\kappa-1}{2\kappa+1}}}{\log(pd_n)}\right) / \log(1/\rho),$$

and the sample size $n \geq C_2(\log p)^{\frac{2\kappa+1}{2\kappa-1}}$ *for sufficiently large* C_2*, we have*

$$\mathcal{E}^{(T^*)} \leq \frac{2C_1R^2}{1-\rho} n^{-\frac{2\kappa-1}{2\kappa+1}} \log(pd_n),$$

with probability at least $1 - C_0(T^*+1)(Rsn^{-\frac{2\kappa}{2\kappa+1}} + 1/p)$*.*

The non-asymptotic estimation error bound (16) consists of two parts: an optimization error which is incurred by the non-convexity and a statistical error which is incurred by the observation noise and the spline approximation. Here, optimization error decays geometrically with the iteration number t, while the statistical error remains the same when t grows. When the tensor covariate is of order-one, i.e., a vector covariate, the overall optimization problem reduces to classical vector nonparametric additive model. In that case, we do not have the optimization error ($\rho^{t+1}\mathcal{E}^{(0)}$) any more since the whole optimization is convex, and the statistical error term matches the state-of-the-art rate in Huang et al. (2010).

Lastly, let us define $\widehat{\mathcal{T}}(\mathcal{X}) = \sum_{h=1}^{d_n} \sum_{r=1}^{R} \langle \boldsymbol{\beta}_{1hr}^{(T^*)} \circ \boldsymbol{\beta}_{2hr}^{(T^*)} \circ \boldsymbol{\beta}_{3hr}^{(T^*)}, \mathcal{F}_h(\mathcal{X}) \rangle$ as a final estimator of the target function $\mathcal{T}^*(\mathcal{X})$. For any function f on $[a,b]$, we define its $\ell_2(P)$ norm by $\|f\|_2 = \sqrt{\int_a^b f^2(x) dP(x)}$. The following corollary provides the final error rate for the estimation of tensor additive nonparametric function $\mathcal{T}^*(\mathcal{X})$. It incorporates the approximation error of nonparametric component incurred by the B-spline series expansion, and the estimation error of unknown tensor parameter incurred by noises.

Corollary 22 *Suppose Conditions 10, 15-17, 19 hold and the number of iterations t as well as the sample size n satisfy the requirement in Theorem 21. Assume the non-zero elements in $\mathcal{T}^*(\cdot)$ are the same as $\sum_{h=1}^{d_h} \langle \mathcal{B}_h^*, \mathcal{F}_h(\cdot) \rangle$. Then, the final estimator satisfies*

$$\left\| \widehat{\mathcal{T}} - \mathcal{T}^* \right\|_2^2 = \mathcal{O}_p\Big(n^{-\frac{2\kappa}{2\kappa+1}} \log(pd_n) \Big).$$

5. Simulations

In this section, we carry out intensive simulation studies to evaluate the performance of our STAR method, and compare it with existing competing solutions including the tensor linear regression (TLR) (Zhou et al., 2013), the Gaussian process based nonparametric method (GP) (Kanagawa et al., 2016), and the nonlinear tensor regression via alternative minimization procedure (AMP) (Suzuki et al., 2016). We find that STAR enjoys better performance both in terms of prediction accuracy and computational efficiency.

Throughout our numerical studies, the natural cubic splines with B-spline basis are used in STAR with the degree fixed to be five, which amounts to having four inner knots. For both STAR and TLR, five-fold cross-validation is employed to select the best pair of the tuning parameters R and λ, where the tensor rank R is chosen from $\{2,3\}$ and λ is selected from a sequence that is uniformly distributed on the logarithm scale in an interval $[10^{-5}, 1]$. For GP and AMP, as suggested by Kanagawa et al. (2016), the Gaussian kernel is used and the bandwidth is set to be 100; five-fold cross-validation is used to select λ, where λ is selected from the same range that is used for TLR and STAR.

5.1 Low-rank covariate structure

The simulated data are generated based on the following model,

$$y_i = \mathcal{T}^*(\mathcal{X}_i) + \sigma\epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_i \sim \mathrm{N}(0,1)$. For each observation, $y_i \in \mathbb{R}$ is the response and $\mathcal{X}_i \in \mathbb{R}^{p_1 \times p_2}$ is the two-way tensor (matrix) covariate. We fix $p_2 = 8$, and we vary n from $\{400, 600\}$, p_1 from $\{20, 50, 100\}$, and σ from $\{0.1, 1\}$. We assume that there are 10 and 4 important features along the first and second way of $\mathcal{X}$, respectively.

Since both GP and AMP models require a low-rank structure on the tensor covariates, in this simulation we consider the low-rank covariate case which favors their models. For each $i = 1, \ldots, n$, we consider $\mathcal{X}_i = \boldsymbol{x}_i^{(1)} \circ \boldsymbol{x}_i^{(2)}$, where the elements of $\boldsymbol{x}_i^{(1)} \in \mathbb{R}^{p_1}$ and $\boldsymbol{x}_i^{(2)} \in \mathbb{R}^{p_2}$ are independently and identically distributed from uniform distribution. Following the additive model that is considered in Ravikumar et al. (2009), we generate samples according

Table 1. *MSE of simulated data with low-rank covariate structure.*

(n, σ)	model	$p_1 = 20$		$p_1 = 50$		$p_1 = 100$	
$(400, 0.1)$	**STAR**	0.51	(0.01)	0.53	(0.01)	0.50	(0.01)
	TLR	2.03	(0.17)	2.63	(0.17)	3.16	(0.48)
	AMP	1.02	(0.02)	1.01	(0.02)	1.01	(0.02)
	GP	1.02	(0.01)	1.03	(0.03)	1.02	(0.01)
$(600, 0.1)$	**STAR**	0.50	(0.01)	0.50	(0.01)	0.52	(0.01)
	TLR	2.11	(0.09)	2.81	(0.14)	3.19	(0.21)
	AMP	0.99	(0.02)	1.01	(0.00)	1.00	(0.02)
	GP	0.99	(0.02)	1.01	(0.01)	1.02	(0.01)
$(400, 1)$	**STAR**	1.54	(0.02)	1.59	(0.03)	1.56	(0.01)
	TLR	2.29	(0.17)	3.18	(0.40)	4.91	(0.56)
	AMP	1.98	(0.02)	2.01	(0.01)	2.06	(0.03)
	GP	2.05	(0.04)	2.04	(0.03)	2.07	(0.03)
$(600, 1)$	**STAR**	1.55	(0.01)	1.53	(0.01)	1.54	(0.01)
	TLR	2.89	(0.39)	3.90	(0.36)	4.69	(0.46)
	AMP	2.02	(0.02)	2.03	(0.03)	2.04	(0.03)
	GP	2.02	(0.01)	2.04	(0.04)	2.06	(0.03)

[*] The simulation compares sparse tensor additive regression (STAR), tensor linear regression (TLR), alternative minimizing procedure (AMP), and Gaussian process (GP). The reported errors are the medians over 20 independent runs with the standard error given in parentheses.

to

$$y_i = \sum_{j=1}^{10}\sum_{k=1}^{4} \mathcal{T}_j^*\left([\boldsymbol{x}_i^{(1)}]_j\right) \cdot \mathcal{T}_k^*\left([\boldsymbol{x}_i^{(2)}]_k\right) + \sigma\epsilon_i,\ i = 1, \ldots, n,$$

where the nonlinear functions $\mathcal{T}_j^*$ and $\mathcal{T}_k^*$ are given by

$$\mathcal{T}_j^*(x) = \begin{cases} -\sin(1.5x), & \text{if } j \text{ is odd,} \\ x^3 + 1.5(x-0.5)^2, & \text{if } j \text{ is even,} \end{cases} \text{ and } \mathcal{T}_k^*(x) = \begin{cases} -\phi(x, 0.5, 0.8^2), & \text{if } k \text{ is odd,} \\ \sin\{\exp(-0.5x)\}, & \text{if } k \text{ is even.} \end{cases}$$

Here $\phi(\cdot, 0.5, 0.8^2)$ is the probability density function of the normal distribution $\mathrm{N}(0.5, 0.8^2)$.

Table 1 compares the mean squared error (MSE) of all four models, where MSE is assessed on an independently generated test data of $2,000$ samples. The STAR model shows the lowest MSE in all cases. As expected, the TLR model has unsatisfactory performance, as the true regression model is non-linear. The two nonparametric tensor regression models GP and AMP can capture partial nonlinear structures, however, our method still outperform theirs.

Next, we investigate the computational costs of all four methods. Table 2 compares the computation time in the example with $n = 400$ and $\sigma = 0.1$. The results of other scenarios are similar and hence omitted. All the computation time includes the model fitting and the model tuning using five-fold cross-validation. Overall our STAR method is as fast as AMP and is much faster than GP. When the tensor dimension is small, $p_1 = 20$, the linear model TLR is the fastest one, however, its computation cost dramatically increase when the dimension p_1 increases, and is even slower than other nonparametric models when

Table 2. *The computation time of all methods in Section 5.1 with $n = 400$ and $\sigma = 0.1$.*

p_1	20	50	100
STAR	353.30	227.65	391.90
TLR	156.19	738.96	1803.49
AMP	330.74	336.10	341.15
GP	1841.51	1913.27	1792.50

[*] All the time include five-fold cross-validation procedures to tune the parameters. The results are averaged over 20 independent runs. The experiment was conducted using a single processor Inter(R) Xeon(R) CPU E5-2600@2.60GHz.

$p_1 = 100$. On the other hand, the computation time of our model is less sensitive to the dimensionality and is even faster than TLR when p_1 is large. This indicates the importance of fully exploiting the low-rankness and sparsity structures in order to improve computational efficiency.

5.2 General covariate structure

The settings are similar as that in Section 5.1, except that the covariate is not low-rank. Here, the elements of $\mathcal{X}_i \in \mathbb{R}^{p_1 \times p_2}$ are independently and identically distributed from uniform distribution. In particular, we generate n observations according to

$$y_i = \sum_{j=1}^{10}\sum_{k=1}^{4} \mathcal{T}_{jk}^*([\mathcal{X}_i]_{jk}) + \sigma\epsilon_i, \ i = 1, \dots, n,$$

where

$$\mathcal{T}_{jk}^*(x) = \begin{cases} -\sin(1.5x), & \text{if } j \text{ is odd and } k \text{ is odd,} \\ x^3 + 1.5(x-0.5)^2, & \text{if } j \text{ is even and } k \text{ is odd,} \\ -\phi(x, 0.5, 0.8^2), & \text{if } j \text{ is odd and } k \text{ is even,} \\ \sin\{\exp(-0.5x)\}, & \text{if } j \text{ is even and } k \text{ is even.} \end{cases}$$

To run GP and AMP models, we perform singular value decomposition in order to meet the requirements of the low-rank tensor inputs.

The comparisons of the MSE are summarized in Table 3: the MSE of the STAR model is much lower than that of all the other three models. Similar to the experiment in Section 5.1, the large MSE of TLR is attributed to the incapability of capturing the nonlinear relationship in the additive model. In this example, the GP and AMP models deliver relatively large MSE because the assumption of low-rank covariate structure is violated. Importantly, the MSE of our STAR model decreases, as the sample size increases or the noise level σ decreases. These observations align with our theoretical finding in Theorem 21.

To test the adaptability of STAR, we further consider a linear data-generating model:

$$y_i = \sum_{j=1}^{10}\sum_{k=1}^{4} \mathcal{T}_{jk}^*([\mathcal{X}_i]_{jk}) + \sigma\epsilon_i, \ i = 1, \dots, n,$$

Table 3. *MSE of simulated data with general covariate structure.*

(n, σ)	model	$p_1 = 20$		$p_1 = 50$		$p_1 = 100$	
$(400, 0.1)$	**STAR**	2.30	(0.06)	3.34	(0.22)	5.04	(0.33)
	TLR	40.62	(0.53)	53.94	(0.71)	92.31	(1.60)
	AMP	41.09	(0.31)	40.78	(0.35)	40.97	(0.30)
	GP	41.62	(0.24)	41.35	(0.38)	41.18	(0.22)
$(600, 0.1)$	**STAR**	1.76	(0.04)	2.14	(0.11)	2.53	(0.12)
	TLR	37.12	(0.63)	43.74	(0.85)	58.75	(0.88)
	AMP	40.39	(0.30)	40.35	(0.35)	41.00	(0.37)
	GP	40.65	(0.66)	40.57	(0.64)	41.28	(0.44)
$(400, 1)$	**STAR**	3.98	(0.11)	5.28	(0.26)	7.17	(0.50)
	TLR	42.12	(0.51)	56.47	(0.78)	94.09	(2.38)
	AMP	42.00	(0.38)	42.40	(0.47)	41.51	(0.28)
	GP	42.56	(0.36)	42.19	(0.37)	42.08	(0.30)
$(600, 1)$	**STAR**	3.13	(0.05)	3.63	(0.13)	4.10	(0.11)
	TLR	38.06	(0.59)	45.43	(0.65)	60.20	(1.03)
	AMP	41.44	(0.54)	41.99	(0.45)	41.76	(0.40)
	GP	41.31	(0.50)	41.87	(0.44)	42.49	(0.53)

[*] The simulation compares sparse tensor additive regression (STAR), tensor linear regression (TLR), alternative minimizing procedure (AMP), and Gaussian process (GP). The reported errors are the medians over 20 independent runs with the standard error given in parentheses.

Table 4. *MSE of simulated data from linear models.*

model	$p_1 = 20$		$p_1 = 50$		$p_1 = 100$	
STAR	1.40	(0.03)	1.62	(0.13)	1.85	(0.06)
TLR	1.19	(0.01)	1.52	(0.02)	2.42	(0.04)
AMP	2.07	(0.02)	2.08	(0.02)	2.09	(0.03)
GP	2.10	(0.01)	2.10	(0.01)	2.10	(0.01)

[*] The simulation compares sparse tensor additive regression (STAR), tensor linear regression (TLR), alternative minimizing procedure (AMP), and Gaussian process (GP). The reported errors are the medians over 20 independent runs with the standard error given in parentheses.

where

$$\mathcal{T}^*_{jk}(x) = \begin{cases} 0.5x, & \text{if } j \text{ is odd,} \\ x, & \text{if } j \text{ is even.} \end{cases}$$

We use $n = 400$ and $\sigma = 1$ for illustration and summarize the result in Table 4. We observe TLR delivers the lowest MSE and STAR is the second best when $p_1 = 20$ and 50. The MSE of STAR is the lowest when $p_1 = 100$. Thus the STAR model is competitive with TLR in general when the data generating model is truly linear.

5.3 Three-way covariate structure

We next extend the previous simulations to a three-way covariate structure. We consider two cases in this section. In the first case, we generate the tensor covariate $\mathcal{X}_i \in \mathbb{R}^{p_1 \times p_2 \times 2}$

Table 5. *MSE of simulated data with three-way tensor covariates.*

		$p_1 = 20$		$p_1 = 50$		$p_1 = 200$	
Case 1							
$(600, 0.1)$	**STAR**	0.34	(0.03)	1.33	(0.23)	1.10	(0.03)
	TLR	10.69	(0.06)	10.86	(0.14)	13.47	(0.25)
$(600, 1)$	**STAR**	1.45	(0.02)	1.71	(0.20)	1.99	(0.01)
	TLR	11.93	(0.09)	12.45	(0.14)	18.14	(0.95)
$(1000, 0.1)$	**STAR**	0.37	(0.02)	0.43	(0.03)	1.24	(0.25)
	TLR	10.82	(0.10)	10.77	(0.12)	11.68	(0.11)
$(1000, 1)$	**STAR**	1.40	(0.03)	1.47	(0.02)	2.02	(0.03)
	TLR	11.90	(0.06)	11.98	(0.15)	13.58	(0.19)
Case 2							
$(600, 0.1)$	**STAR**	0.78	(0.01)	2.07	(1.10)	1.01	(0.00)
	TLR	13.11	(0.11)	13.11	(0.17)	16.62	(0.33)
$(600, 1)$	**STAR**	2.07	(0.08)	2.01	(0.02)	1.99	(0.02)
	TLR	14.45	(0.17)	14.33	(0.14)	20.73	(0.66)
$(1000, 0.1)$	**STAR**	0.75	(0.00)	0.78	(0.05)	1.35	(0.17)
	TLR	13.15	(0.22)	12.98	(0.17)	13.52	(0.23)
$(1000, 1)$	**STAR**	1.76	(0.02)	2.13	(0.12)	1.99	(0.02)
	TLR	14.15	(0.29)	14.05	(0.19)	15.57	(0.27)

* The simulation compares sparse tensor additive regression (STAR) and tensor linear regression (TLR). The reported errors are the medians over 20 independent runs, and the standard error of the medians are given in parentheses. All the methods use five-fold cross-validation procedures to tune the parameters.

whose elements are from i.i.d. uniform distribution, and then generate the response $y_i \in \mathbb{R}$ from

$$\text{case 1: } y_i = \sum_{j=1}^{10} \sum_{k=1}^{4} \left[\sin(\mathcal{T}_{jk1}^{*}([\mathcal{X}_i]_{jk1})) + \log |\mathcal{T}_{jk2}^{*}([\mathcal{X}_i]_{jk2})| \right] + \sigma \epsilon_i, \ i = 1, \ldots, n,$$

where $\mathcal{T}_{jkl}^{*}$ with $l = 1, 2$ is defined the same as the expression (**??**). In the second case, we generate the response from

$$\text{case 2: } y_i = \sin\left(\sum_{j=1}^{10} \sum_{k=1}^{4} \mathcal{T}_{jk}^{*}([\mathcal{X}_i]_{jkl}) \right) + \log \left| \sum_{j=1}^{10} \sum_{k=1}^{4} \mathcal{T}_{jk}^{*}([\mathcal{X}_i]_{jkl}) \right| + \sigma \epsilon_i, \ i = 1, \ldots, n.$$

It is worth noting that case 1 uses an additive model while case 2 does not. Therefore, the additive model assumption in our STAR method is actually mis-specified in case 2.

Since the softwares of AMP and GP models for three-way tensor covariates are not available, in this simulation we compare our STAR model only with TLR. We vary the sample size $n \in \{600, 1000\}$, the first-way dimension $p_1 \in \{20, 50, 200\}$, the noise level $\sigma \in \{0.1, 1\}$, and fix the second-way dimension $p_2 = 10$. Similar to previous simulations, we assume that there are 10, 4, and 2 important features along the three modes of the tensor $\mathcal{X}_i$, respectively. As shown in Table 5, the MSE of the STAR model is consistently lower than that of TLR, even in the case when the additive model is mis-specified.

6. An Application to Online Advertising

In this section, we apply the STAR model to click-through rate (CTR) prediction in online advertising. The CTR is defined to be the ratio between the number of clicks and the number of impressions (ad views). In this study, we are interested in predicting the overall CTR, which is the average CTR across different ad campaigns. The overall CTR is an effective measure to evaluate the performance of online advertising. A low overall CTR usually indicates that the ads are not effectively displayed or the wrong audience is being targeted. As a reference, the across-industry overall CTR of display campaigns in the United States from April 2016 to April 2017 is 0.08%.[1] Importantly, the CTR is also closely related to the revenue. Define the effective revenue per mile (eRPM) to be the amount of revenue from every 1000 impressions, and we have eRPM $= 1000 \times$ CPC $\times$ CTR, where CPC is the cost per click. From this expression, we can see that a good CTR prediction is critical to ad pricing, and the CTR prediction is a highly important task in online advertising.

We collect 136 ad campaign data during 28 days from a premium Internet media company.[2] The data from each day have been aggregated into six time periods and each of the 136 campaigns involves ads delivered via three devices: phone, tablet, and personal computer (PC). In total, we have $224 = 28 \times 8$ time periods. There are 153 million of users in total, and we divide all the users into two groups, a younger group and an elder group, which are partitioned by the median age. For each time period, we aggregate the number of impressions of 136 advertising campaigns that are delivered on each of three types of devices for each of the two age groups. Denoting the number of impressions by $\mathcal{X}$, each data point has $\mathcal{X}_i \in \mathbb{R}^{136\times 3\times 2}$, and $i = 1, 2, \ldots, 224$ represents the time period. In this study, we aim to study the relationship between the overall CTR and the three-way tensor covariate of impressions.

Figure 1 delineates the marginal relationship between the overall CTR and the impression of one advertisement that is delivered on phone, tablet, and PC, respectively. In this example, the overall CTR clearly reveals a non-linear pattern across all devices. Moreover, it is generally believed that not all ads have significant impacts on the overall CTR and hence ad selection is an active research area (Choi et al., 2010; Xu et al., 2016). To fulfill both tasks of capturing the nonlinear relationship and selecting important ads, we apply the proposed STAR model to predict the overall CTR. The logarithm transformations are applied to both the CTR and the number of impressions. We train and tune each method on the data obtained on the first 24 days, and use the remaining data as the test data to assess the prediction accuracy. The MSE of our STAR model is 0.51, which is much lower than 5.44, the MSE of the TLR. This result shows the effectiveness of capturing the non-linear relationship as well as assuming the low-rankness and group sparsity structures both in increasing the CTR prediction accuracy and the algorithm efficiency. The AMP and GP models are not compared due to the lack of implementation for three-way covariates.

In terms of ad selection, the STAR model with group lasso penalty selects 60 out of 136 ads, as well as all three devices and two age groups as active variables for the CTR prediction. As a comparison, TLR selects 114 ads, 46 of which are also selected by our STAR

[1] The data are from `http://www.richmediagallery.com/learn/benchmarks`.

[2] The reported data and results in this section are deliberately incomplete and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

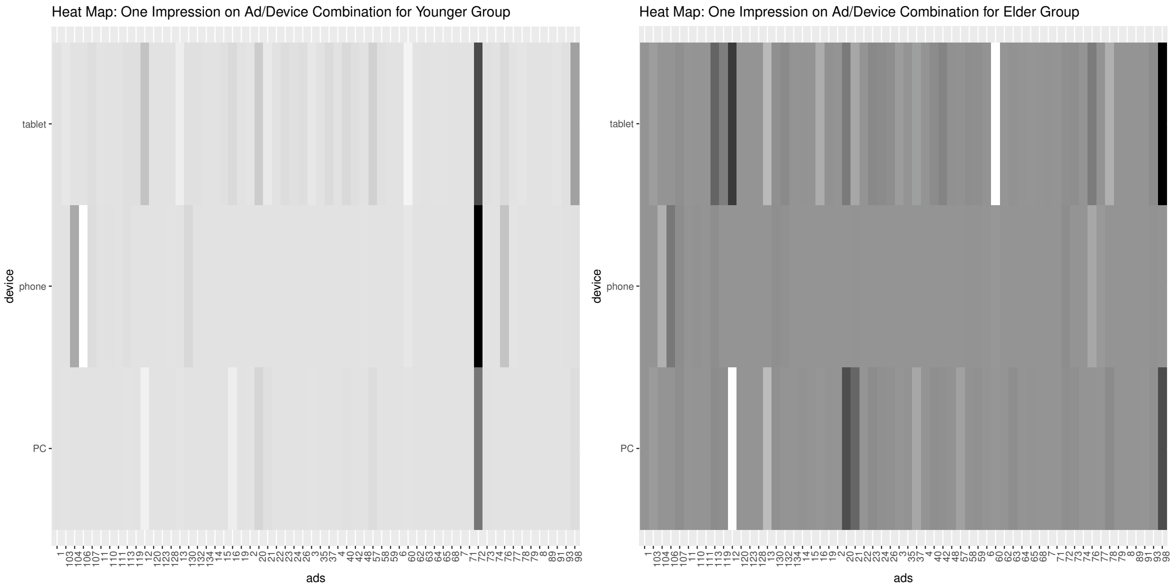

Figure 4. *Heatmaps for the overall CTR. The left panel is the mean change in the overall CTR if the test data have one additional impression on each ad and device combination for the younger group and the right panel is for the elder group. Darker tiles indicate greater positive mean change in the overall CTR and lighter tiles indicate greater negative mean change. The IDs of ads have been renumbered for concerns of confidentiality.*

method. Besides the prediction on the overall CTR and the ad selection performance, we are also able to see which combination of ad, device, and age group yields the most significant impact on the overall CTR. In the left panel of Figure 4, each tile represents a combination of ad and device for the younger group, and the darkness of the tile implies the sensitivity of the overall CTR associated with one more impression on this combination; the right panel of Figure 4 shows the heatmap for the elder group. Displayed on phones of the elder users, the ad with ID 98 has the most positive effect on the overall CTR. Figure 5 is plotted similarly except that the change is due to every 1000 additional impressions on the certain combinations. The overall CTR has the largest growth when 1000 additional impressions are allocated to the ad with ID 73 displaying on phones of the younger users. The different patterns between Figure 4 and 5 indicate the nonlinear relationship between the overall CTR and the number of impressions. This result is important for managerial decision making. Under a specific budget, our STAR model facilitates ad placement targeting based on the best ad/device/age combination to maximize the ad revenue.

Appendix

In this appendix, we present the detailed proofs for Theorem 13, 21 and Corollary 22.

A1 Proof of Theorem 13

First, we state a key lemma which quantifies the estimation error of each component individually within one iteration step.

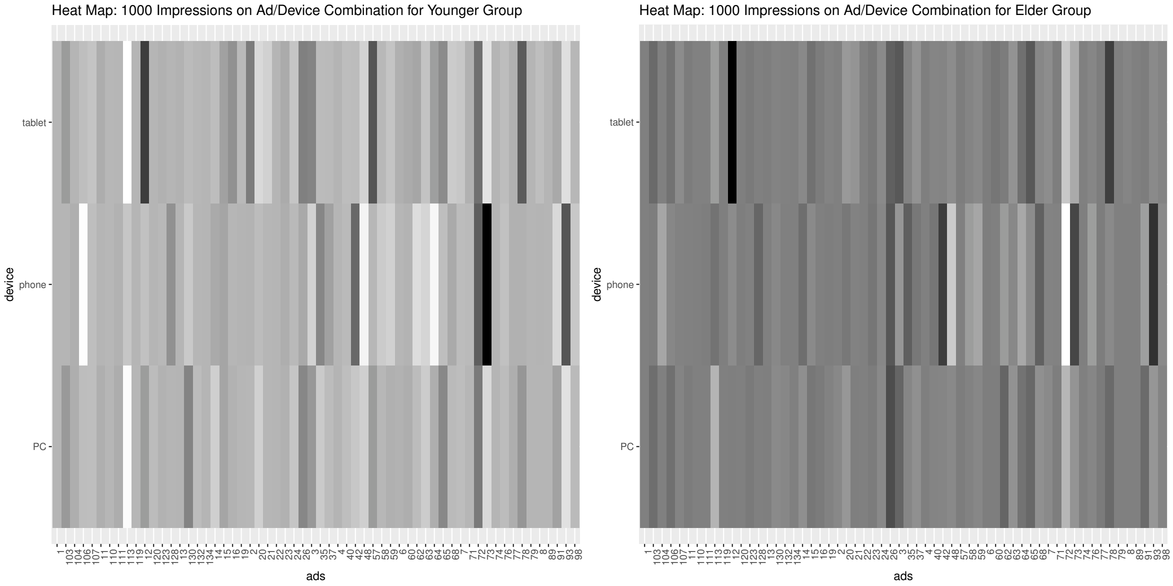

Figure 5. *Heatmaps for the overall CTR. The left panel is the mean change in the overall CTR if the test data have 1000 additional impressions on each ad and device combination for the younger group and the right panel is for the elder group. Darker tiles indicate greater positive mean change in the overall CTR and lighter tiles indicate greater negative mean change. The IDs of ads have been renumbered for concerns of confidentiality.*

Lemma 23 *Suppose Conditions 4-6 and 8 hold, and the updates at time t satisfy $\boldsymbol{b}_2^{(t)} \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3^{(t)} \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$. Let the penalty $\mathcal{P}$ fulfills the decomposable property (See Definition 7 for details), and the regularization parameter $\lambda_{1n}^{(t)} \geq 4\varepsilon_1 + (\mu_{2n}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 + \mu_{3n}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2)/\Phi(\mathcal{S}_1)$ where μ_{2n}, μ_{3n} and ε_1 are defined in Condition 4 and 8 respectively. Then the update of $\boldsymbol{b}_1$ at time $t+1$ satisfies*

$$\|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2 \leq 4\lambda_{1n}^{(t)}\Phi(\mathcal{S}_1)/\gamma_{1n}, \tag{A17}$$

with probability at least $1 - (\delta_1 + \delta_2 + \delta_3)$, where γ_{1n} is defined in Condition 6 and $\Phi(\mathcal{S}_1)$ is the support space compatibility constant defined in (13)*.*

Proof. For notation simplicity, we will drop the superscript of $\boldsymbol{b}_1^{(t)}, \boldsymbol{b}_2^{(t)}, \boldsymbol{b}_3^{(t)}, \lambda_{1n}^{(t)}$ and replace the superscript of $\boldsymbol{b}_1^{(t+1)}, \boldsymbol{b}_2^{(t+1)}, \boldsymbol{b}_3^{(t+1)}$ by $\boldsymbol{b}_1^+, \boldsymbol{b}_2^+, \boldsymbol{b}_3^+$ in the rest of the proof for Lemma 23.

First of all, the loss function (10) enjoys a bi-convex structure, in the sense that $\mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3)$ is convex in one argument when fixing the other two. Then, given current update $\boldsymbol{b}_2, \boldsymbol{b}_3$, the penalized alternating minimization with respect to $\boldsymbol{b}_1$ takes the form of

$$\boldsymbol{b}_1^+ = \underset{\boldsymbol{b}_1}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) + \lambda_{1n}\mathcal{P}(\boldsymbol{b}_1).$$

As $\boldsymbol{b}_1^+$ minimizes the loss function, we have

$$\mathcal{L}(\boldsymbol{b}_1^+, \boldsymbol{b}_2, \boldsymbol{b}_3) + \lambda_{1n}\mathcal{P}(\boldsymbol{b}_1^+) \leq \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) + \lambda_{1n}\mathcal{P}(\boldsymbol{b}_1^*), \tag{A18}$$

which further implies the following inequality by the convexity of $\mathcal{L}(\cdot, \boldsymbol{b}_2, \boldsymbol{b}_3)$,

$$\begin{aligned} \lambda_{1n}(\mathcal{P}(\boldsymbol{b}_1^+) - \mathcal{P}(\boldsymbol{b}_1^*)) &\leq \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1^+, \boldsymbol{b}_2, \boldsymbol{b}_3) \\ &\leq \underbrace{|\langle \nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle|}_{\text{RHS}}. \end{aligned} \tag{A19}$$

Recall that $\nabla_1 \mathcal{L}$ is the noisy gradient function with respect to $\boldsymbol{b}_1$ defined in (12). To separate the statistical error and optimization error, we utilize noiseless gradient function $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)$ defined in (12) as a bridge. The detail decomposition is presented as follows,

$$\begin{aligned} \text{RHS} \quad \leq \quad & \underbrace{|\langle \nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle|}_{\text{statistical error}} \\ & + \underbrace{|\langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle|}_{\text{optimization error}}, \end{aligned}$$

where $\nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*) = 0$. Moreover, based on the decomposability of penalty $\mathcal{P}$ (See Condition 12),

$$\begin{aligned} \text{RHS} \quad \leq \quad & \|\nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\|_{\mathcal{P}^*} \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_{\mathcal{P}} \\ & + \langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle \\ & + \langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle, \end{aligned}$$

where $\mathcal{P}^*$ is the dual norm of $\mathcal{P}$. We write $\mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) = \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_{\mathcal{P}}$. In addition, putting (A19) and Conditions 4 and 8 together, we have

$$\begin{aligned} & |\langle \nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^+ - \boldsymbol{b}_1^* \rangle| \\ & \leq \varepsilon_1 \mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) + \big(\mu_{2n} \|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2 + \mu_{3n} \|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2\big) \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2, \end{aligned} \tag{A20}$$

with probability at least $1 - (\delta_1 + \delta_3)$. Together with (A18),

$$\lambda_{1n}(\mathcal{P}(\boldsymbol{b}_1^+) - \mathcal{P}(\boldsymbol{b}_1^*)) \leq \varepsilon_1 \mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) + \big(\mu_{2n} \|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2 + \mu_{3n} \|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2\big) \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2.$$

Since $\lambda_{1n} \geq 4\varepsilon_1 + \big(\mu_{2n} \|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2 + \mu_{3n} \|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2\big)/\Phi(\mathcal{S}_1)$, we have

$$\mathcal{P}(\boldsymbol{b}_1^+) - \mathcal{P}(\boldsymbol{b}_1^*) \leq \frac{1}{4} \mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) + \Phi(\mathcal{S}_1) \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2. \tag{A21}$$

Again, using the decomposability of $\mathcal{P}$, the LHS of (A21) can be decomposed by

$$\begin{aligned} \mathcal{P}(\boldsymbol{b}_1^+) - \mathcal{P}(\boldsymbol{b}_1^*) \quad &= \quad \mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^* + \boldsymbol{b}_1^*) - \mathcal{P}(\boldsymbol{b}_1^*) \\ &= \quad \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp} + \boldsymbol{b}_1^* + (\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) - \mathcal{P}(\boldsymbol{b}_1^*) \\ &\geq \quad \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp}) + \mathcal{P}(\boldsymbol{b}_1^* + (\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) - \mathcal{P}(\boldsymbol{b}_1^*) \\ &\geq \quad \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp}) - \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}), \end{aligned} \tag{A22}$$

where $\mathcal{S}_1^\perp$ is the complement set of $\mathcal{S}_1$. Equipped with (A21),

$$3\mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp}) \leq 5\mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) + 4\Phi(\mathcal{S}_1) \|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2. \tag{A23}$$

By the definition of support space compatibility constant (13),

$$\mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) \le \Phi(\mathcal{S}_1)\|(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}\|_2 \le \Phi(\mathcal{S}_1)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2.$$

Together with $\mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) \le \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) + \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp})$ and (A23), we obtain

$$\mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) \le 4\Phi(\mathcal{S}_1)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2. \tag{A24}$$

On the other hand, based on sparse strongly convex Condition 6,

$$\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1^+, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^* - \boldsymbol{b}_1^+ \rangle \le -\frac{\gamma_{1n}}{2}\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2^2.$$

with probability at least $1-\delta_2$. Plugging in (A20), we obtain with probability at least $1-(\delta_1+\delta_2+\delta_3)$,

$$\begin{aligned}\frac{\gamma_{1n}}{2}\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2^2 &\le \langle \nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^* - \boldsymbol{b}_1^+ \rangle + \mathcal{L}(\boldsymbol{b}_1^+, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\\ &\le \varepsilon_1 \mathcal{P}(\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*) + \big(\mu_{2n}\|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2 + \mu_{3n}\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2\big)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2 + \lambda_{1n}(\mathcal{P}(\boldsymbol{b}_1^*) - \mathcal{P}(\boldsymbol{b}_1^+)).\end{aligned} \tag{A25}$$

From (A22),

$$\begin{aligned}\lambda_{1n}(\mathcal{P}(\boldsymbol{b}_1^*) - \mathcal{P}(\boldsymbol{b}_1^+)) &\le \lambda_{1n}\Big(\mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}) - \mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1^\perp})\Big)\\ &\le \lambda_{1n}\mathcal{P}((\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*)_{\mathcal{S}_1}).\end{aligned}$$

Together with (A24) and (A25),

$$\begin{aligned}\frac{\gamma_{1n}}{2}\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2^2 &\le \lambda_{1n}\Phi(\mathcal{S}_1)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2 + 4\varepsilon_1\Phi(\mathcal{S}_1)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2\\ &\quad + \big(\mu_{2n}\|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2 + \mu_{3n}\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2\big)\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2.\end{aligned}$$

Dividing by $\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2$ in both sides and plugging in the lower bound of λ_{1n}, it yields that

$$\|\boldsymbol{b}_1^+ - \boldsymbol{b}_1^*\|_2 \le \frac{4\lambda_{1n}\Phi(\mathcal{S}_1)}{\gamma_{1n}},$$

with probability at least $1-(\delta_1+\delta_2+\delta_3)$. This ends the proof. ■

Note that (A17) is a generic result since we have not provided a detail form for certain parameters. Similar results also hold for the update of $\boldsymbol{b}_2^{(t)}, \boldsymbol{b}_3^{(t)}$ (see next corollary) and detailed proofs are omitted here.

Corollary 24 *Suppose Conditions 29-34 hold, and the updates at time t satisfy $\boldsymbol{b}_1^{(t)} \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*)$, $\boldsymbol{b}_2^{(t)} \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3^{(t)} \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$. With the regularization parameters $\lambda_{2n}^{(t)}, \lambda_{3n}^{(t)}$ satisfy*

$$\begin{aligned}\lambda_{2n}^{(t)} &\ge 4\varepsilon_2 + (\mu'_{1n}\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2 + \mu'_{3n}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2)/\Phi(\mathcal{S}_2)\\ \lambda_{3n}^{(t)} &\ge 4\varepsilon_3 + (\mu''_{1n}\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_2^*\|_2 + \mu''_{2n}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_3^*\|_2)/\Phi(\mathcal{S}_3)\end{aligned}$$

and penalty $\mathcal{P}$ fulfills the decomposable property, then the updates of $\boldsymbol{b}_2, \boldsymbol{b}_3$ at time $t+1$ satisfy

$$\begin{aligned}\|\boldsymbol{b}_2^{(t+1)} - \boldsymbol{b}_2^*\|_2 &\le 4\lambda_{2n}^{(t)}\Phi(\mathcal{S}_2)/\gamma_{2n}\\ \|\boldsymbol{b}_3^{(t+1)} - \boldsymbol{b}_3^*\|_2 &\le 4\lambda_{3n}^{(t)}\Phi(\mathcal{S}_3)/\gamma_{3n},\end{aligned}$$

with probability at least $1-(\delta_1+\delta_2+\delta_3)$.

Now we are ready to prove the main theorem. Applying the result in Lemma 23, and plugging in the lower bound of $\lambda_{1n}^{(t)}$, we have

$$\|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2 \le \frac{4\mu_{2n}}{\gamma_{1n}}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 + \frac{4\mu_{3n}}{\gamma_{1n}}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2 + \frac{16\varepsilon_1\Phi(\mathcal{S}_1)}{\gamma_{1n}}.$$

Taking the square in both sides and noticing that $(a+b+c)^2 \le 3(a^2+b^2+c^2)$,

$$\|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2^2 \le 3\Big(\frac{4\mu_{2n}}{\gamma_{1n}}\Big)^2\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2^2 + 3\Big(\frac{4\mu_{3n}}{\gamma_{1n}}\Big)^2\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2^2 + 3\Big(\frac{16\varepsilon_1\Phi(\mathcal{S}_1)}{\gamma_{1n}}\Big)^2,$$

with probability at least $1-(\delta_1+\delta_2+\delta_3)$. Similarly, applying Corollary 24, we have

$$\begin{aligned}\|\boldsymbol{b}_2^{(t+1)} - \boldsymbol{b}_2^*\|_2^2 &\le 3\Big(\frac{4\mu'_{1n}}{\gamma_{2n}}\Big)^2\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2^2 + 3\Big(\frac{4\mu'_{3n}}{\gamma_{2n}}\Big)^2\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2^2 + 3\Big(\frac{16\varepsilon_2\Phi(\mathcal{S}_2)}{\gamma_{2n}}\Big)^2\\ \|\boldsymbol{b}_3^{(t+1)} - \boldsymbol{b}_3^*\|_2^2 &\le 3\Big(\frac{4\mu''_{1n}}{\gamma_{3n}}\Big)^2\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2^2 + 3\Big(\frac{4\mu''_{2n}}{\gamma_{3n}}\Big)^2\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2^2 + 3\Big(\frac{16\varepsilon_3\Phi(\mathcal{S}_3)}{\gamma_{3n}}\Big)^2,\end{aligned}$$

with probability at least $1-(\delta_1+\delta_2+\delta_3)$. Denote $\mathcal{E}^{(t+1)} = \|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2^2 + \|\boldsymbol{b}_2^{(t+1)} - \boldsymbol{b}_2^*\|_2^2 + \|\boldsymbol{b}_3^{(t+1)} - \boldsymbol{b}_3^*\|_2^2$. Adding the above three bounds together, it implies

$$\begin{aligned}\mathcal{E}^{(t+1)} \quad \le \quad & 48\Big(\Big[\frac{\mu_{1n}^{'2}}{\gamma_{2n}^2} + \frac{\mu_{1n}^{''2}}{\gamma_{3n}^2}\Big]\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2^2 + \Big[\frac{\mu_{2n}^2}{\gamma_{1n}^2} + \frac{\mu_{2n}^{''2}}{\gamma_{3n}^2}\Big]\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2^2 + \Big[\frac{\mu_{3n}^2}{\gamma_{1n}^2} + \frac{\mu_{3n}^{'2}}{\gamma_{2n}^2}\Big]\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2^2\Big)\\ & +768\Big(\frac{\varepsilon_1^2\Phi(\mathcal{S}_1)^2}{\gamma_{1n}^2} + \frac{\varepsilon_2^2\Phi(\mathcal{S}_2)^2}{\gamma_{2n}^2} + \frac{\varepsilon_3^2\Phi(\mathcal{S}_3)^2}{\gamma_{3n}^2}\Big).\end{aligned}$$

Define the contraction parameter

$$\rho = 288\max\{\mu_{1n}^{'2}, \mu_{1n}^{''2}, \mu_{2n}^2, \mu_{2n}^{''2}, \mu_{3n}^2, \mu_{3n}^{'2}\}/\min\{\gamma_{1n}^2, \gamma_{2n}^2, \gamma_{3n}^2\},$$

then

$$\mathcal{E}^{(t+1)} \le \rho\mathcal{E}^{(t)} + C_0\Big(\frac{\varepsilon_1^2\Phi(\mathcal{S}_1)^2}{\gamma_{1n}^2} + \frac{\varepsilon_2^2\Phi(\mathcal{S}_2)^2}{\gamma_{2n}^2} + \frac{\varepsilon_3^2\Phi(\mathcal{S}_3)^2}{\gamma_{3n}^2}\Big),$$

with probability at least $1-3(\delta_1+\delta_2+\delta_3)$. This ends the proof. ■

A2 Proof of Theorem 21

Moreover, let $\alpha = \min\{\alpha_1, \alpha_2, \alpha_3\}$, $p = \max\{p_1, p_2, p_3\}$ and $s = \max\{s_1, s_2, s_3\}$, where s_i is the cardinality of $\mathcal{S}_i$ defined in (7).

Our proof consists of three steps. First, we verify Conditions 4-6 and 8 in Lemma 25-27 for B-spline basis function and give explicit forms of Lipschitz-gradient parameter, sparse-strongly-convex parameter and statistical error. Second, we prove a generic contraction result by the induction argument. Last, we combine results in first two steps and achieve the final estimation rate.

At first, Lemma 25 and 26 show that the loss function in (8) with B-spline basis function is sparse strongly convex and Lipschitz continuous. The proofs are deferred to Sections 3.1 and 3.2.

Lemma 25 *Consider $\{\psi_{jklh}(x)\}_{h=1}^{d_n}$ introduced in* (3) *are normalized B-spline basis functions and suppose Conditions Conditions 15-16 and 17 hold. When $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_1^*)$, $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$, the loss function $\mathcal{L}(\cdot,\cdot,\cdot)$ is sparse strongly convex in its first argument, namely*

$$\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_1 \mathcal{L}\big(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3), \boldsymbol{b}_1^* - \boldsymbol{b}_1 \rangle \le -\frac{\gamma_{1n}}{2} \|\boldsymbol{b}_1 - \boldsymbol{b}_1^*\|_2^2, \tag{A26}$$

where $\gamma_{1n} = C_1(1+o(1))Rd_n^{-1}s^2c_*^4$.

Lemma 26 *Suppose $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*), \boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$ and Conditions Conditions 15-16 and 17 hold. Considering the B-spline basis function, we have with probability at least $1 - 12/p$,*

$$T_1 = \big\langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\rangle \le \mu_{2n} \big\|\boldsymbol{b}_1 - \boldsymbol{b}_1^*\big\|_2 \big\|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\big\|_2$$
$$T_2 = \big\langle \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3^*), \boldsymbol{b}_1 - \boldsymbol{b}_1^* \big\rangle \le \mu_{3n} \big\|\boldsymbol{b}_1 - \boldsymbol{b}_1^*\big\|_2 \big\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\big\|_2,$$

where $\mu_{2n} = \mu_{3n} = 12(s^3R^2/d_n^2 + C_0\sqrt{\log p/n})R^2s^2c^{*4}$.

The verification of Conditions 31-34 and derivation of $\gamma_{2n}, \gamma_{3n}, \mu'_{1n}, \mu'_{3n}, \mu''_{1n}, \mu''_{2n}$ remain the same and only differ in some constants. Thus, we let

$$\begin{aligned} &\max\{\mu_{2n}, \mu_{3n}, \mu'_{1n}, \mu'_{3n}, \mu''_{1n}, \mu''_{2n}\} = C_3(s^3R^2/d_n^2 + \sqrt{\log p/n})R^2s^2c^{*4} \\ &\min\{\gamma_{1n}, \gamma_{2n}, \gamma_{3n}\} = C_4(1+o(1))Rd_n^{-1}s^2c_*^4 \end{aligned} \tag{A27}$$

for some absolute constant C_3, C_4.

Next lemma gives an explicit bound on statistical error for the update of $\boldsymbol{b}_1$ when we utilize B-spline basis and choose the penalty $\mathcal{P}$ to be group lasso penalty.

Lemma 27 *Suppose Conditions 15-16 and 17 hold and Consider $\{\psi_{jklh}(x)\}_{h=1}^{d_n}$ introduced in* (3) *to be normalized B-spline basis function. For $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$, we have with probability at least $1 - C_0Rd_ns/n$,*

$$\begin{aligned} &\big\|\nabla_1 \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1 \widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\big\|_{\mathcal{P}^*} \\ \le\ & C_1Rc^{*4}\Big(\frac{s^5}{d_n^{\kappa-1/2}}\sqrt{\frac{\log(en)}{n}} + \frac{s^6}{d_n^{\kappa+1/2}} + \sigma\sqrt{\frac{s^4\log(pd_n)}{n}}\Big). \end{aligned}$$

for some absolute constants C_0, C_1, where $0 < \kappa < 1$ describes the smoothness of function class $\mathcal{H}$ defined in (15).

We complete the proof of Theorem 21 by the induction argument. When $t = 1$, the initialization condition naturally holds by Condition 19. Suppose $\|\boldsymbol{b}_1^{(t)} - \boldsymbol{b}_1^*\|_2 \le \alpha$, $\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 \le \alpha$, $\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2 \le \alpha$ holds for some $t \ge 1$. For $t = t+1$, first we utilize the result in Lemma 23 and plug in the lower bound of $\lambda_{1n}^{(t)}$,

$$\begin{aligned}
\|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2 &\le \frac{4\lambda_{1n}^{(t)}\Phi(\mathcal{S}_1)}{\gamma_{1n}} \\
&\le \frac{4\Phi(\mathcal{S}_1)}{\gamma_{1n}}\Big(4\varepsilon_1 + \big(\mu_{2n}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 + \mu_{3n}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2\big)/\Phi(\mathcal{S}_1)\Big) \\
&\le \frac{16\Phi(\mathcal{S}_1)\varepsilon_1}{\gamma_{1n}} + \frac{4\mu_{2n}}{\gamma_{1n}}\|\boldsymbol{b}_2^{(t)} - \boldsymbol{b}_2^*\|_2 + \frac{4\mu_{3n}}{\gamma_{1n}}\|\boldsymbol{b}_3^{(t)} - \boldsymbol{b}_3^*\|_2 \\
&\le \frac{16\Phi(\mathcal{S}_1)\varepsilon_1}{\gamma_{1n}} + \frac{4}{\gamma_{1n}}\Big(\mu_{2n}\alpha + \mu_{3n}\alpha\Big).
\end{aligned}$$

As long as the statistical error ε_1 satisfies

$$\varepsilon_1 \le \Big(1 - \frac{4(\mu_{2n} + \mu_{3n})}{\gamma_{1n}}\Big)\frac{\alpha\gamma_{1n}}{40\Phi(\mathcal{S}_1)}, \tag{A28}$$

we have $\|\boldsymbol{b}_1^{(t+1)} - \boldsymbol{b}_1^*\|_2 \le \alpha$. The proofs for $\|\boldsymbol{b}_2^{(t+1)} - \boldsymbol{b}_2^*\|_2 \le \alpha$ and $\|\boldsymbol{b}_3^{(t+1)} - \boldsymbol{b}_3^*\|_2 \le \alpha$ are similar when $\varepsilon_2, \varepsilon_3$ satisfy

$$\begin{aligned}
\varepsilon_2 &\le \Big(1 - \frac{4(\mu_{1n} + \mu_{3n})}{\gamma_{2n}}\Big)\frac{\alpha\gamma_{2n}}{16\Phi(\mathcal{S}_2)}, \qquad \text{(A29)}\\
\varepsilon_3 &\le \Big(1 - \frac{4(\mu_{1n} + \mu_{2n})}{\gamma_{3n}}\Big)\frac{\alpha\gamma_{3n}}{16\Phi(\mathcal{S}_3)}.
\end{aligned}$$

Second, when $\boldsymbol{b}_2, \boldsymbol{b}_3$ are fixed, the update scheme for $\boldsymbol{b}_1$ exactly fits the one in Huang et al. (2010) with group lasso penalty under B-spline basis function expansion. Define $\mathcal{S}_1^{(t)} = \{j \in [p_1] | \|\boldsymbol{\beta}_{1j}^{(t)}\|_2 \neq 0\}$. Similar to the proof of first part in Theorem 1 in Huang et al. (2010), we could obtain $|\mathcal{S}_1^{(t)}| \le C_0|\mathcal{S}_1| = C_0 s$ for a finite constant $C_0 > 1$ with probability converging to 1. That means the number of non-zero elements in the estimator from group-lasso-type penalization is comparable with the size of true support. The guarantee for $\boldsymbol{b}_2^{(t)}, \boldsymbol{b}_3^{(t)}$ remains the same.

Therefore, we can conclude that $\boldsymbol{b}_1^{(t)} \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_1^*), \boldsymbol{b}_2^{(t)} \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*), \boldsymbol{b}_3^{(t)} \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$ hold for any iteration $t = 1, 2, \dots$ as long as the statistical error is sufficiently small such that (A28)-(A29) hold. We choose the tuning parameter $\lambda_{1n}^{(t)}, \lambda_{2n}^{(t)}, \lambda_{3n}^{(t)}$ as defined in Condition 10 with generic parameters specified in Lemma 26, 27. Repeatedly applying the result in Theorem 13 and summing from $t = 1$ to $t = t+1$, one can provide a generic form of error updates,

$$\mathcal{E}^{(t+1)} \le \rho^{t+1}\mathcal{E}^{(0)} + \frac{1 - \rho^{t+1}}{1 - \rho}C_0\Big(\frac{\varepsilon_1^2\Phi(\mathcal{S}_1)^2}{\gamma_{1n}^2} + \frac{\varepsilon_2^2\Phi(\mathcal{S}_2)^2}{\gamma_{2n}^2} + \frac{\varepsilon_3^2\Phi(\mathcal{S}_3)^2}{\gamma_{3n}^2}\Big), \tag{A30}$$

with probability at least $1 - 2(t+1)(\delta_1 + \delta_2 + \delta_3)$. As before, (A30) still provides a generic form of error updates.

Finally, we combine results from Lemmas 25-27. According to (A27), the contraction parameter is upper bounded by

$$\rho \leq \frac{288C_3^2c^{*8}}{C_4^2c_*^8}\Big(\frac{s^6}{d_n^2}+\frac{d_n^2\log p}{n}\Big).$$

When the sample size n is large enough such that

$$d_n^2 \geq \frac{1}{4}\frac{288C_3^2c^{*8}s^6}{C_4^2c_*^8},\ n \geq \frac{1}{4}\frac{288C_3^2d_n^2(\log p)c^{*8}}{c_*^8C_4^2}, \tag{A31}$$

one can guarantee $\rho \leq 1/2$. For group lasso penalty (9), it has been shown in Wainwright (2014) that $\max\{\Phi(\mathcal{S}_1),\Phi(\mathcal{S}_2),\Phi(\mathcal{S}_3)\} = s$. Then, we can have an explicit form for the upper bound in (A30),

$$\begin{aligned}
\mathcal{E}^{(t+1)} &\leq \rho^{t+1}\mathcal{E}^{(0)}+\frac{1-\rho^{t+1}}{1-\rho}3C_0\frac{\max(\varepsilon_1^2,\varepsilon_2^2,\varepsilon_3^2)\max(\Phi(\mathcal{S}_1)^2,\Phi(\mathcal{S}_2)^2,\Phi(\mathcal{S}_3)^2)}{\min(\gamma_{1n}^2,\gamma_{2n}^2,\gamma_{3n}^2)}\\
&\leq \rho^{t+1}\mathcal{E}^{(0)}+\frac{1-\rho^{t+1}}{1-\rho}3\frac{C_0R^2}{(1+o(1))}\Big(\frac{s^2}{d_n^{-2}s^4c_*^8}\Big)c^{*8}3\Big(\frac{s^{10}}{d_n^{2\kappa-1}}\frac{\log ep}{n}+\frac{s^{12}}{d_n^{2\kappa+1}}+\frac{\sigma^2\log(pd_n)}{n}\Big)\\
&= \rho^{t+1}\mathcal{E}^{(0)}+\frac{1-\rho^{t+1}}{1-\rho}\frac{9C_0c^{*8}R^2}{c_*^8(1+o(1))}\Big(\frac{s^8}{d_n^{2\kappa-3}}\frac{\log ep}{n}+\frac{s^{10}}{d_n^{2\kappa-1}}+\sigma^2\frac{d_n^2}{s^2}\frac{\log(pd_n)}{n}\Big),
\end{aligned} \tag{A32}$$

with probability at least $1-C_0(t+1)(Rd_ns/n+1/p)$. From Conditions 16-17, we known that s,σ,c_*,c^* are all bounded by some absolute constants. Then (A32) can be further simplified as

$$\mathcal{E}^{(t+1)} \leq \rho^{t+1}\mathcal{E}^{(0)}+\frac{C_1R^2}{(1+o(1))}\frac{1-\rho^{t+1}}{1-\rho}\Big(\frac{\log ep}{d_n^{2\kappa-3}n}+\frac{1}{d_n^{2\kappa-1}}+\sigma^2\frac{d_n^2\log(pd_n)}{s^2n}\Big).$$

To trade-off the statistical error part ($\sigma^2d_n^2\frac{\log pd_n}{s^2n}$) and approximation error part ($\frac{\log ep}{d_n^{2\kappa-3}n}+\frac{1}{d_n^{2\kappa-1}}$), one can take $d_n \asymp n^{\frac{1}{2\kappa+1}}$. Then the above bound will reduce to

$$\mathcal{E}^{(t+1)} \leq \rho^{t+1}\mathcal{E}^{(0)}+\frac{C_1R^2}{(1-\rho)(1+o(1))}n^{-\frac{2\kappa-1}{2\kappa+1}}\log(pd_n),$$

with proper adjustments for the constant C_1. Moreover, when the total number of iterations is no smaller than

$$T^* = \log\Big(\frac{(1-\rho)(1+o(1))}{C_1\mathcal{E}^{(0)}}\frac{n^{\frac{2\kappa-1}{2\kappa+1}}}{\log(pd_n)}\Big)/\log(1/\rho),$$

we have with probability at least $1-C_0(T^*+1)(Rsn^{-\frac{2\kappa}{2\kappa+2}}+1/p)$,

$$\mathcal{E}^{(T^*)} \leq \frac{2C_1R^2}{(1-\rho)(1+o(1))}n^{-\frac{2\kappa-1}{2\kappa+1}}\log(pd_n),$$

as long as $n \geq C_2(\log p)^{\frac{2\kappa+1}{2\kappa-1}}$ for sufficiently large C_2. This sample complexity is sufficient to guarantee that (A28)-(A29) and (A31) hold under Conditions 16-17. This ends the proof. ■

A3 Proof of Corollary 22

Recall that $\widetilde{\mathcal{T}}(\mathcal{X}) = \sum_{h=1}^{d_n} \langle \mathcal{B}_h^*, \mathcal{F}_h(\mathcal{X}) \rangle$, where $\mathcal{B}_h^* = \sum_{r=1}^{R} \boldsymbol{\beta}_{1hr}^* \circ \boldsymbol{\beta}_{2hr}^* \circ \boldsymbol{\beta}_{3hr}^*$, and $\widehat{\mathcal{T}}(\mathcal{X}) = \sum_{h=1}^{d_n} \langle \widehat{\mathcal{B}}_h, \mathcal{F}_h(\mathcal{X}) \rangle$, where $\widehat{\mathcal{B}}_h = \sum_{r=1}^{R} \boldsymbol{\beta}_{1hr}^{(T^*)} \circ \boldsymbol{\beta}_{2hr}^{(T^*)} \circ \boldsymbol{\beta}_{3hr}^{(T^*)}$. We make the following decomposition,

$$\left\| \widehat{\mathcal{T}} - \mathcal{T}^* \right\|_2^2 \le 2 \underbrace{\left\| \widehat{\mathcal{T}} - \widetilde{\mathcal{T}} \right\|_2^2}_{I_1} + 2 \underbrace{\left\| \widetilde{\mathcal{T}} - \mathcal{T}^* \right\|_2^2}_{I_2},$$

where $\|f\|_2 = \sqrt{\int_a^b f^2(x) dP(x)}$. Intuitively, I_1 quantifies the estimation error of $\{\mathcal{B}_h^*\}_{h=1}^{d_n}$, while I_2 measures the overall approximation error by using B-spline basis function expansion. We bound I_1 and I_2 in two steps.

1. By the definition, it's easy to see

$$\sum_{h=1}^{d_n} \left\| \widehat{\mathcal{B}}_h - \mathcal{B}_h^* \right\|_F^2 \le 3R \sum_{r=1}^{R} \sum_{h=1}^{d_n} \left(\|\boldsymbol{\beta}_{1hr}^{(T^*)} - \boldsymbol{\beta}_{1hr}^*\|_2^2 + \|\boldsymbol{\beta}_{2hr}^{(T^*)} - \boldsymbol{\beta}_{2hr}^*\|_2^2 + \|\boldsymbol{\beta}_{3hr}^{(T^*)} - \boldsymbol{\beta}_{3hr}^*\|_2^2 \right).$$

According to the basis property of spline expansions (De Boor et al., 1978), we reach that

$$\begin{aligned} I_1 &\le C_1 d_n^{-1} \sum_{h=1}^{d_n} \left\| \widehat{\mathcal{B}}_h - \mathcal{B}_h^* \right\|_F^2 \\ &\le 3C_1 R d_n^{-1} \sum_{r=1}^{R} \sum_{h=1}^{d_n} \left(\|\boldsymbol{\beta}_{1hr}^{(T^*)} - \boldsymbol{\beta}_{1hr}^*\|_2^2 + \|\boldsymbol{\beta}_{2hr}^{(T^*)} - \boldsymbol{\beta}_{2hr}^*\|_2^2 + \|\boldsymbol{\beta}_{3hr}^{(T^*)} - \boldsymbol{\beta}_{3hr}^*\|_2^2 \right) \\ &= 3C_1 R d_n^{-1} \mathcal{E}^{(T^*)}. \end{aligned}$$

According to Theorem 21,

$$I_1 \le 3C_1 R d_n^{-1} n^{-\frac{2\kappa-1}{2\kappa+1}} \log p d_n, \tag{A33}$$

with probability at least $1 - C_0(T^* + 1)(s n^{-\frac{2\kappa}{2\kappa+1}} + 1/p)$.

2. By the assumption of CP-low-rankness, we have

$$\begin{aligned} I_2 &= \Big\| \sum_{h=1}^{d_n} \langle \mathcal{B}_h^*, \mathcal{F}_h(\mathcal{X}) \rangle - \mathcal{T}^*(\mathcal{X}) \Big\|_2^2 \\ &= \Big\| \sum_{j=1}^{p_1} \sum_{k=1}^{p_2} \sum_{l=1}^{p_3} (f_{jkl}^{d_n}(\mathcal{X}_{jkl}) - f_{jkl}^*(\mathcal{X}_{jkl})) \Big\|_2^2. \end{aligned}$$

According to Lemma 28, we have

$$I_2 \le C_2 s^6 d_n^{-2\kappa}. \tag{A34}$$

Putting (A33) and (A34) together, we reach

$$\left\|\widehat{\mathcal{T}}-\mathcal{T}^*\right\|_2^2 \leq 3C_1 R d_n^{-1} n^{-\frac{2\kappa-1}{2\kappa+1}} \log p d_n + C_2 s^6 d_n^{-2\kappa}.$$

Note that under Condition 15-19, both R and s are bounded. By taking $d_n \asymp n^{-\frac{1}{2\kappa+1}}$, we have

$$\left\|\widehat{\mathcal{T}}-\mathcal{T}^*\right\|_2^2 = \mathcal{O}_p\Big(n^{-\frac{2\kappa}{2\kappa+1}} \log p d_n\Big).$$

This ends the proof. ■

Acknowledgments

The authors thank the action editor Francis Bach and two reviewers for their helpful comments and suggestions which led to a much improved presentation. Will Wei Sun acknowledges support from the Office of Naval Research (ONR N00014-18-1-2759). Jingfei Zhang acknowledges support from the National Science Foundation (NSF DMS-2015190). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, or the Office of Naval Research.

References

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1): 77–120, 2017.

Eric C Chi and Tamara G. Kolda. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299, 2012.

Yejin Choi, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Mauricio Mediano, and Bo Pang. Using landing pages for sponsored search ad selection. In *WWW*, 2010.

Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763, 2017.

Weiwei Guo, Irene Kotsia, and Ioannis Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2012.

Botao Hao, Will Wei Sun, Yufeng Liu, and Guang Cheng. Simultaneous clustering and estimation of heterogeneous graphical models. *The Journal of Machine Learning Research*, 18(1):7981–8038, 2017.

Botao Hao, Anru Zhang, and Guang Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *arXiv preprint arXiv:1801.09326*, 2018.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3):1169, 2015.

Jian Huang, Joel L Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282, 2010.

Hung Hung and Chen-Chien Wang. Matrix variate logistic regression model with application to eeg data. *Biostatistics*, 14(1):189–202, 2012.

Heishiro Kanagawa, Taiji Suzuki, Hayato Kobayashi, Nobuyuki Shimizu, and Yukihiro Tagami. Gaussian process nonparametric tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, pages 1632–1641, 2016.

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

Bing Li, Min Kyung Kim, Naomi Altman, et al. On dimension folding of matrix-or array-valued statistical objects. *The Annals of Statistics*, 38(2):1094–1121, 2010.

Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Lukas Meier, Sara Van de Geer, Peter Bühlmann, et al. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097, 2011.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 11 2012. doi: 10.1214/12-STS400.

Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 21–28. ACM, 2009.

Guillaume Rabusseau and Hachem Kadri. Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, 2016.

Garvesh Raskutti, Ming Yuan, Han Chen, et al. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71 (5):1009–1030, 2009. ISSN 1467-9868.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, pages 689–705, 1985.

Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

Will Wei Sun and Lexin Li. Dynamic tensor clustering. *Journal of the American Statistical Association*, 114(528):1708–1725, 2019.

Will Wei Sun, Junwei Lu, Han Liu, and Guang Cheng. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):899–916, 2017.

Taiji Suzuki, Heishiro Kanagawa, Hayato Kobayashi, Nobuyuki Shimizu, and Yukihiro Tagami. Minimax optimal alternating minimization for kernel nonparametric tensor learning. In *Advances in Neural Information Processing Systems*, pages 3783–3791, 2016.

Martin J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1(1):233–253, 2014. doi: 10.1146/annurev-statistics-022513-115643.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Jian Xu, Xuhui Shao, Jianjie Ma, Kuang chih Lee, Hang Qi, and Quan Lu. Lift-based bidding in ad selection. In *AAAI*, 2016.

Xinyang Yi and Constantine Caramanis. Regularized em algorithms: A unified framework and statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

Rose Yu and Yan Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, 2016.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

Ming Yuan and Cunhui Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16:1031–1068, 2016.

Anru Zhang. Cross: Efficient low-rank tensor completion. *Annals of Statistics*, 47:936–964, 2019.

Anru Zhang and Rungang Han. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*, 114:1708–1725, 2019.

Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang chih Lee, Zenglin Xu, Yuan Qi, and Zoubin Ghahramani. Distributed flexible nonlinear tensor factorization. In *Advances in Neural Information Processing Systems*, 2016.

Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108:540–552, 2013.

S Zhou, X Shen, and DA Wolfe. Local asymptotics for regression splines and confidence regions. *The Annals of statistics*, 26(5):1760–1782, 1998.

Hongtu Zhu, Yasheng Chen, Joseph G. Ibrahim, Yimei Li, Colin Hall, and Weili Lin. Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association*, 104(487):1203–1212, 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.tm08096.

Supplementary Materials
Sparse Tensor Additive Regression

Botao Hao, Boxiang Wang, Pengyuan Wang,
Jingfei Zhang, Jian Yang, Will Wei Sun

In the supplementary, we present the definition and properties of the B-spline basis, some additional conditions for our theoretical results and the detailed proofs of Lemmas 25-27.

1. Properties of B-spline

We formally define the q-th order B-splines with a set of m internal knot sequences $k = \{0 = k_0 < k_1 < \ldots < k_m < k_{m+1} = 1\}$ recursively,

$$b_l^1(x) = \begin{cases} 1, k_l \leq x < k_{l+1} \\ 0, \text{otherwise} \end{cases}$$

and

$$b_l^q(x) = \frac{x - k_l}{k_{l+q-1} - k_l} b_l^{q-1}(x) + \frac{k_{l+q} - x}{k_{l+q} - k_{l+1}} b_{l+1}^{q-1}. \tag{A1}$$

Then under some smoothness conditions, $f(x) \approx s(x) = \sum_l b_l^q(x)\beta_l = \boldsymbol{b}(x)^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}_i \in \mathbb{R}^p$ with $p = m + q$. For the random variable X satisfying Condition 15, we have $\mathbb{E}[b_l^q(X)] \leq C_1 d_n^{-1}, \mathbb{E}[b_l^q(X)]^2 \leq C_2 d_n^{-1}$ for some constants C_1 and C_2. The detailed proofs refer to Stone (1985); Huang et al. (2010); Fan et al. (2011).

Additionally, we restate the result in Huang et al. (2010) for the approximation error rate under B-spline basis function.

Lemma 28 (Stone (1985); Huang et al. (2010)) *Suppose Condition 15 holds and if the number of spline series is chosen by $d_n = \mathcal{O}(n^{1/(2\kappa+1)})$. Then there exists an $f_{jkl}^{d_n} \in \mathcal{S}_n$ satisfying*

$$\left\| f_{jkl}^{d_n} - f_{jkl}^* \right\|_2^2 = \mathcal{O}_p(d_n^{-2\kappa}) = \mathcal{O}_p(n^{-2\kappa/(2\kappa+1)}). \tag{A2}$$

2. Additional conditions for Section 4.1

In this section, we present addition conditions for Lipschitz-gradient (Conditions 29-30), sparse strongly convex (Conditions 31-32), and statistical error (Conditions 33-34) for the update of $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$. We define $\nabla_2\mathcal{L}(\cdot,\cdot,\cdot)$ and $\nabla_3\mathcal{L}(\cdot,\cdot,\cdot)$ are the gradient taken with respect to the second and the third argument.

Condition 29 () *For $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*), \boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$, the noiseless gradient function $\nabla_2\widetilde{\mathcal{L}}(\cdot, \boldsymbol{b}_2^*, \boldsymbol{b}_3)$ satisfies μ_{1n}'-Lipschitz-gradient condition, and $\nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \cdot)$ satisfies μ_{3n}'-Lipschitz-gradient condition. That is,*

$$\begin{aligned} \left\langle \nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3) - \nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3), \boldsymbol{b}_2 - \boldsymbol{b}_2^* \right\rangle &\leq \mu_{1n}' \left\| \boldsymbol{b}_2 - \boldsymbol{b}_2^* \right\|_2 \left\| \boldsymbol{b}_1 - \boldsymbol{b}_1^* \right\|_2 \\ \left\langle \nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3) - \nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_2 - \boldsymbol{b}_2^* \right\rangle &\leq \mu_{3n}' \left\| \boldsymbol{b}_2 - \boldsymbol{b}_2^* \right\|_2 \left\| \boldsymbol{b}_3 - \boldsymbol{b}_3^* \right\|_2, \end{aligned}$$

with probability at least $1 - \delta_1$.

Condition 30 () *For* $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*), \boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$*, the noiseless gradient function* $\nabla_3\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \cdot, \boldsymbol{b}_3^*)$ *satisfies* $\mu_{2n}^{''}$*-Lipschitz-gradient condition, and* $\nabla_3\widetilde{\mathcal{L}}(\cdot, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*)$ *satisfies* $\mu_{1n}^{''}$*-Lipschitz-gradient condition. That is,*

$$\begin{aligned}\big\langle \nabla_3\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - \nabla_3\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_3 - \boldsymbol{b}_3^* \big\rangle &\le \mu_{2n}^{''} \big\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\big\|_2 \big\|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\big\|_2 \\ \big\langle \nabla_3\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*) - \nabla_3\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2^*, \boldsymbol{b}_3^*), \boldsymbol{b}_3 - \boldsymbol{b}_3^* \big\rangle &\le \mu_{1n}^{''} \big\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\big\|_2 \big\|\boldsymbol{b}_1 - \boldsymbol{b}_1^*\big\|_2,\end{aligned}$$

with probability at least $1 - \delta_1$*.*

Condition 31 () *For any* $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*), \boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$*, the loss function* $\mathcal{L}(\cdot,\cdot,\cdot)$ *is sparse strongly convex in its first variable, namely*

$$\mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_2\mathcal{L}\big(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3), \boldsymbol{b}_2^* - \boldsymbol{b}_2 \rangle \ge \frac{\gamma_{2n}}{2}\|\boldsymbol{b}_2 - \boldsymbol{b}_2^*\|_2^2,$$

with probability at least $1 - \delta_2$*. Here,* $\gamma_{2n} > 0$ *is the strongly convex parameter.*

Condition 32 () *For any* $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*), \boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$*, the loss function* $\mathcal{L}(\cdot,\cdot,\cdot)$ *is sparse strongly convex in its first variable, namely*

$$\mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_3\mathcal{L}\big(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3^*), \boldsymbol{b}_3 * -\boldsymbol{b}_3 \rangle \ge \frac{\gamma_{3n}}{2}\|\boldsymbol{b}_3 - \boldsymbol{b}_3^*\|_2^2,$$

with probability at least $1 - \delta_2$*. Here,* $\gamma_{3n} > 0$ *is the strongly convex parameter.*

Condition 33 () *For any* $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*)$*,* $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha_3,s_3}(\boldsymbol{b}_3^*)$*, we have with probability at least* $1 - \delta_3$*,*

$$\big\|\nabla_2\mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3) - \nabla_2\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2^*, \boldsymbol{b}_3)\big\|_{\mathcal{P}^*} \le \varepsilon_2.$$

Condition 34 () *For any* $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha_1,s_1}(\boldsymbol{b}_1^*)$*,* $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha_2,s_2}(\boldsymbol{b}_2^*)$*, we have with probability at least* $1 - \delta_3$*,*

$$\big\|\nabla_3\mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3^*) - 3\widetilde{\mathcal{L}}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3^*)\big\|_{\mathcal{P}^*} \le \varepsilon_3.$$

3. Proofs of Lemmas 25-27

In this section, we present the proof of Lemmas 25-27. If X is sub-Gaussian random variable, then its ϕ_2-Orlicz norm can be bounded such that $\|X\|_{\phi_2} \le C_1$ for some absolute constant. If X is sub-exponential random variable, then its ϕ_1-Orlicz norm can be bounded such that $\|X\|_{\phi_1} \le C_2$ for some absolute constant C_2.

3.1 Proof of Lemma 25

Recall that $\boldsymbol{b}_1 = (\vartheta_{11}^\top, \dots, \vartheta_{1p_1}^\top)^\top \in \mathbb{R}^{Rd_np_1\times 1}$. Define $\mathcal{S}_1' = \{j \in [p] | \|\vartheta_{1j}\|_2 \neq 0 \cup \|\vartheta_{1j}^*\|_2 \neq 0\}$, $\boldsymbol{F}_{\mathcal{S}_1'}^1 = (\boldsymbol{F}_j^1 \in \mathbb{R}^{n\times Rd_n}, j \in \mathcal{S}_1')$, $\boldsymbol{b}_{1\mathcal{S}_1'} = (\boldsymbol{\beta}_{1j} \in \mathbb{R}^{Rd_n\times 1}, j \in \mathcal{S}_1')$. Since $\boldsymbol{b}_1 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_1^*)$, we know that $|\mathcal{S}_1'| = C_0 s$ for some positive constant $C_0 \ge 1$ not depending on s. Without loss of generality, assume $|\mathcal{S}_1'| = \{1, \cdots, C_0 s\}$. First, we do some simplifications for the left side of (A26). According to the derivation in (12), we have

$$\begin{aligned}&\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) \\ =\ & \frac{1}{n}\Big(\boldsymbol{b}_1^{*\top}\boldsymbol{F}^{1\top}\boldsymbol{F}^1\boldsymbol{b}_1^* - \boldsymbol{b}_1^\top\boldsymbol{F}^{1\top}\boldsymbol{F}^1\boldsymbol{b}_1 - 2\boldsymbol{y}^\top\boldsymbol{F}^1\boldsymbol{b}_1^* + 2\boldsymbol{y}^\top\boldsymbol{F}^1\boldsymbol{b}_1\Big) \\ =\ & \frac{1}{n}\Big(\boldsymbol{b}_{1\mathcal{S}_1'}^{*\top}\boldsymbol{F}_{\mathcal{S}_1'}^{1\top}\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}^* - \boldsymbol{b}_{1\mathcal{S}_1'}^\top\boldsymbol{F}_{\mathcal{S}_1'}^{1\top}\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'} - 2\boldsymbol{y}^\top\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}^* + 2\boldsymbol{y}^\top\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}\Big),\end{aligned}$$

and

$$
\begin{aligned}
& \langle \nabla_1 \mathcal{L}\big(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3\big), \boldsymbol{b}_1^* - \boldsymbol{b}_1 \rangle \\
= \; & \frac{2}{n}\Big(\boldsymbol{b}_1^{*\top}\boldsymbol{F}^{1\top}\boldsymbol{F}^1\boldsymbol{b}_1^* - \boldsymbol{b}_1^{*\top}\boldsymbol{F}^{1\top}\boldsymbol{F}^1\boldsymbol{b}_1 - \boldsymbol{y}^\top\boldsymbol{F}^1\boldsymbol{b}_1^* + \boldsymbol{y}^\top\boldsymbol{F}^1\boldsymbol{b}_1\Big) \\
= \; & \frac{2}{n}\Big(\boldsymbol{b}_{1\mathcal{S}_1'}^{*\top}\boldsymbol{F}_{\mathcal{S}_1'}^{1\top}\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}^* - \boldsymbol{b}_{1\mathcal{S}_1'}^{*\top}\boldsymbol{F}_{\mathcal{S}_1'}^{1\top}\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'} - \boldsymbol{y}^\top\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}^* + \boldsymbol{y}^\top\boldsymbol{F}_{\mathcal{S}_1'}^1\boldsymbol{b}_{1\mathcal{S}_1'}\Big).
\end{aligned}
$$

Putting the above two equations together, we reach

$$
\begin{aligned}
& \mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \mathcal{L}(\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3) - \langle \nabla_1 \mathcal{L}\big(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3\big), \boldsymbol{b}_1^* - \boldsymbol{b}_1 \rangle \\
= \; & (\boldsymbol{b}_{1\mathcal{S}_1'} - \boldsymbol{b}_{1\mathcal{S}_1'}^*)^\top\Big(- \frac{\boldsymbol{F}_{\mathcal{S}_1'}^{1\top}\boldsymbol{F}_{\mathcal{S}_1'}^1}{n}\Big)(\boldsymbol{b}_{1\mathcal{S}_1'} - \boldsymbol{b}_{1\mathcal{S}_1'}^*).
\end{aligned}
$$

It remains to prove

$$
(\boldsymbol{b}_{1S_1} - \boldsymbol{b}_{1S_1}^*)^\top\Big(\frac{\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1}{n}\Big)(\boldsymbol{b}_{1S_1} - \boldsymbol{b}_{1S_1}^*) \geq \frac{\gamma_{1n}}{2}\|\boldsymbol{b}_{1S_1} - \boldsymbol{b}_{1S_1}^*\|_2^2.
$$

If one can show that $\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1/n \succeq \widetilde{m}\boldsymbol{I}_{C_0 s}$ i.e. the minimal eigenvalue $\sigma_{\min}(\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1/n) \geq \widetilde{m}$ for some positive $\widetilde{m} \in \mathbb{R}$, then we have the strongly convex parameter $\gamma_{1n} = \widetilde{m}$. Let $\boldsymbol{a} = (\boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_{C_0 s}^\top)^\top$ where $\boldsymbol{a}_j = (\boldsymbol{a}_{j1}^\top, \ldots, \boldsymbol{a}_{jR}^\top)^\top \in \mathbb{R}^{Rd_n \times 1}$ and $\boldsymbol{a}_{jr} \in \mathbb{R}^{d_n \times 1}$. Our proof consists of two steps.

Step One. Consider a single coordinate $\boldsymbol{F}_j^1$. For $k \in [p]$ and $j \in [p]$, define

$$
Z_{jkl} = \begin{pmatrix} \psi_{jkl1}([\mathcal{X}_1]_{jkl}) & \cdots & \psi_{jkld_n}([\mathcal{X}_1]_{jkl}) \\ \vdots & \ddots & \vdots \\ \psi_{jkl1}([\mathcal{X}_n]_{jkl}) & \cdots & \psi_{jkld_n}([\mathcal{X}_n]_{jkl}) \end{pmatrix} \in \mathbb{R}^{n \times d_n},
$$

$$
\boldsymbol{D}_{klr} = \begin{pmatrix} \beta_{21rk}\beta_{31rl} & & \\ & \ddots & \\ & & \beta_{2d_nrk}\beta_{3d_nrl} \end{pmatrix} \in \mathbb{R}^{d_n \times d_n}.
$$

By using the triangle inequality and Lemma 3 in Stone (1985), we have for $j \in [C_0 s]$,

$$
C_1 \sum_{k=1}^p \sum_{l=1}^p \sum_{r=1}^R \Big\| Z_{jkl}\boldsymbol{D}_{klr}\boldsymbol{a}_{jr} \Big\|_2^2 \leq \|\boldsymbol{F}_j^1 \boldsymbol{a}_j\|_2^2, \tag{A3}
$$

where C_1 is some positive constant. Divided by n in both sides, we have

$$
C_1 \sum_{k=1}^p \sum_{l=1}^p \sum_{r=1}^R \boldsymbol{a}_{jr}^\top \boldsymbol{D}_{klr}^\top \frac{Z_{jkl}^\top Z_{jkl}}{n} \boldsymbol{D}_{klr}\boldsymbol{a}_{jr} \leq \boldsymbol{a}_j^\top \frac{\boldsymbol{F}_j^{1\top}\boldsymbol{F}_j^1}{n}\boldsymbol{a}_j.
$$

According to Lemma 6.2 in Zhou et al. (1998), there exists certain constants C_2 and C_3 such that

$$
C_2(1 + o(1))d_n^{-1} \leq \sigma_{\min}\Big(\frac{Z_{jkl}^\top Z_{jkl}}{n}\Big) \leq \sigma_{\max}\Big(\frac{Z_{jkl}^\top Z_{jkl}}{n}\Big) \leq C_3(1 + o(1))d_n^{-1}. \tag{A4}
$$

holds for any k, l. Since $\sigma_{\min}(\boldsymbol{A}\boldsymbol{B}) \geq \sigma_{\min}(\boldsymbol{A})\sigma_{\min}(\boldsymbol{B})$, we can bound the minimum eigenvalue of the weighted B-spline design matrix,

$$\sigma_{\min}\Big(\boldsymbol{D}_{klr}^{\top}\frac{Z_{jkl}^{\top}Z_{jkl}}{n}\boldsymbol{D}_{klr}\Big) \geq C_2(1+o(1))d_n^{-1}(\min_h \beta_{2hk}\beta_{3hl})^2.$$

This will enable us to bound the smallest eigenvalue of $\boldsymbol{F}_j^{1\top}\boldsymbol{F}_j^1/n$ as follows,

$$\begin{aligned}
\boldsymbol{a}_j^{\top}\frac{\boldsymbol{F}_j^{1\top}\boldsymbol{F}_j/n}{\|\boldsymbol{a}_j\|_2^2}\boldsymbol{a}_j &\geq C_1\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{r=1}^{R}\boldsymbol{a}_{jr}^{\top}\frac{(\boldsymbol{D}_{klr}^{\top}Z_{jkl}^{\top}Z_{jkl}\boldsymbol{D}_{klr})/n}{\|\boldsymbol{a}_{jr}\|_2^2}\boldsymbol{a}_{jr}\\
&\geq C_1C_2(1+o(1))d_n^{-1}\sum_{r=1}^{R}\min_h\big(\sum_{k=1}^{p}\beta_{2hrk}^{*2}-\alpha^2\big)\min_h\big(\sum_{l=1}^{p}\beta_{3hrl}^{*2}-\alpha^2\big)\\
&\geq C_1C_2(1+o(1))Rd_n^{-1}(sc_*^2-\alpha^2)^2\\
&\geq \frac{1}{4}C_1C_2(1+o(1))Rd_n^{-1}s^2c_*^4,
\end{aligned}$$

where last inequality is due to $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$ for $\alpha \leq c_*\sqrt{s/2}$ and Condition 17. Let $C_1 = C_1/4$. Therefore, for every $j \in [C_0 s]$,

$$\sigma_{\min}\Big(\frac{\boldsymbol{F}_j^{1\top}\boldsymbol{F}_j^1}{n}\Big) \geq C_1C_2(1+o(1))Rd_n^{-1}s^2c_*^4. \tag{A5}$$

Step Two. By the triangle inequality,

$$C_4\Big(\sum_{j=1}^{C_0 s}\|\boldsymbol{F}_j^1\boldsymbol{b}_j\|_2^2\Big) \leq \|\boldsymbol{F}_{S_1}^1\boldsymbol{b}\|_2^2 = \boldsymbol{b}^{\top}\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1\boldsymbol{b},$$

for some constant C_4, which implies

$$\boldsymbol{a}^{\top}\frac{\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1/n}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a} \geq C_4\Big(\frac{\sum_{j=1}^{C_0 s}\|\boldsymbol{F}_j^1\boldsymbol{a}_j\|_2^2}{n\|\boldsymbol{a}\|_2^2}\Big).$$

Together with (A5), we have

$$\boldsymbol{a}^{\top}\frac{\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}^1/n}{\|\boldsymbol{a}\|_2^2}\boldsymbol{a} \geq C_1C_2C_4(1+o(1))d_n^{-1}s^2c_*^4$$

holds for any $\boldsymbol{a}$. Setting $C_1 = C_1C_2C_4$, it essentially implies

$$\sigma_{\min}\big(\frac{1}{n}\boldsymbol{F}_{S_1}^{1\top}\boldsymbol{F}_{S_1}\big) \geq C_1(1+o(1))Rd_n^{-1}s^2c_*^4,$$

for some constant C_1. We can say the sparse strong convexity holds with $\gamma_{1n} = C_1(1+o(1))Rd_n^{-1}s^2c_*^4$. ■

3.2 Proof of Lemma 26

For notation simplicity, we denote

$$g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)=\sum_{h=1}^{d_n}\langle\mathcal{F}_h(\mathcal{X}_i),\sum_{r=1}^{R}\boldsymbol{\beta}_{1hr}\circ\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr}\rangle,$$

where $\mathcal{F}_h(\mathcal{X}_i)$ is defined in (4). According to the definition of the gradient function, we can rewrite the following inner product as

$$\begin{aligned}&\big\langle\nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3),\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*}\big\rangle\\ =\;&\frac{2}{n}\sum_{i=1}^{n}\Big[g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)-g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3^{*})g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)\Big].\end{aligned}\tag{A6}$$

We will bound T_1 first. The bound for T_2 remains similar. Let's decompose T_1 by three parts,

$$\begin{aligned}T_1\;=\;&\big\langle\nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3^{*})-\nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3^{*}),\boldsymbol{b}_1-\boldsymbol{b}_1^{*}\big\rangle\\ =\;&\frac{2}{n}\sum_{i=1}^{n}\Big[g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)-g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3^{*})g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)\\ &-g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3)+g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3^{*})g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3)\Big]\\ =\;&\underbrace{\frac{2}{n}\sum_{i=1}^{n}\Big[g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2-\boldsymbol{b}_2^{*},\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2,\boldsymbol{b}_3)\Big]}_{T_{11}}\\ &\underbrace{+\frac{2}{n}\sum_{i=1}^{n}\Big[g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2-\boldsymbol{b}_2^{*},\boldsymbol{b}_3)\Big]}_{T_{12}}\\ &\underbrace{-\frac{2}{n}\sum_{i=1}^{n}\Big[g_i(\boldsymbol{b}_1^{*},\boldsymbol{b}_2^{*},\boldsymbol{b}_3^{*})g_i(\boldsymbol{b}_1^{+}-\boldsymbol{b}_1^{*},\boldsymbol{b}_2-\boldsymbol{b}_2^{*},\boldsymbol{b}_3)\Big]}_{T_{13}}.\end{aligned}$$

By writing explicitly of $g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)$,

$$\begin{aligned}g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)\;&=\;\sum_{h=1}^{d_n}\Big(\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}[\mathcal{F}_h(\mathcal{X}_i)_{jkl}]\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}\Big)\\ &=\;\sum_{h=1}^{d_n}\Big(\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\psi_{jklh}([\mathcal{X}_i]_{jkl})\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}\Big).\end{aligned}$$

Since $\sup_x|\psi_{jklh}(x)|\le 1$, the ϕ_2-Orlicz norm for each individual component can be bounded by

$$\Big\|\psi_{jklh}([\mathcal{X}_i]_{jkl})\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}\Big\|_{\phi_2}\le|\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}|,\text{ for }j,k,l\in[p].$$

Based on rotation invariance, we obtain

$$\Big\|g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)\Big\|_{\phi_2} \le \Big(\sum_{h=1}^{d_n}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}(\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl})^2\Big)^{\frac{1}{2}}.$$

In the following, we will bound the expectation of $g_i(\boldsymbol{b}_1^*,\boldsymbol{b}_2-\boldsymbol{b}_2^*,\boldsymbol{b}_3)g_i(\boldsymbol{b}_1^+-\boldsymbol{b}_1^*,\boldsymbol{b}_2,\boldsymbol{b}_3)$. By the property of B-spline basis function (See Section 1) and Cathy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\big(g_i(\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3)\big) &= \sum_{h=1}^{d_n}\Big(\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\mathbb{E}[\mathcal{F}_h(\mathcal{X}_i)_{jkl}]\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}\Big)\\
&\le \frac{1}{d_n}\sum_{h=1}^{d_n}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{r=1}^{R}\beta_{1hrj}\beta_{2hrk}\beta_{3hrl}\\
&\le \frac{s^{\frac{3}{2}}R}{d_n}\Big(\sum_{h=1}^{d_n}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{r=1}^{R}\beta_{1hrj}^2\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

Combining the above ingredients together with Hoeffding's inequality (See Lemma 35), we obtain with probability at least $1-1/p$,

$$\begin{aligned}
T_{11} &\le 2\Big[\frac{s^3R^2}{d_n^2}+C_0\sqrt{\frac{\log p}{n}}\Big]\Big(\sum_{h=1}^{d_n}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{r=1}^{R}\beta_{1hrj}^{*2}(\beta_{2hrk}-\beta_{2hrk}^*)^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}\\
&\quad\times\Big(\sum_{h=1}^{d_n}\sum_{j=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{r=1}^{R}(\beta_{1hrj}^+-\beta_{1hrj}^*)\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

Noting that $\boldsymbol{b}_2\in\mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*),\boldsymbol{b}_3\in\mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$, we have

$$\begin{aligned}
T_{11} &\le 2\Big[\frac{s^3R^2}{d_n^2}+C_0\sqrt{\frac{\log p}{n}}\Big]\max_h(\sum_{j=1}^{p}\beta_{1hrj}^{*2})^{\frac{1}{2}}\max_h(\sum_{k=1}^{p}\beta_{2hrk}^{*2})^{\frac{1}{2}}\max_h(\sum_{l=1}^{p}\beta_{3hrl}^{*2})\big\|\boldsymbol{b}_2-\boldsymbol{b}_2^*\big\|_2\big\|\boldsymbol{b}_1^+-\boldsymbol{b}_1^*\big\|_2\\
&\le 2\Big[\frac{s^3R^2}{d_n^2}+C_0\sqrt{\frac{\log p}{n}}\Big]s^2R^2c^{*4}\big\|\boldsymbol{b}_2-\boldsymbol{b}_2^*\big\|_2\big\|\boldsymbol{b}_1^+-\boldsymbol{b}_1^*\big\|_2,
\end{aligned}$$

where the last inequality is from Condition 17. The upper bounds for T_{12} and T_{13} are similar. Putting them together, with probability at least $1-6/p$,

$$|T_1|\le 6\Big[\frac{s^3R^2}{d_n^2}+C_0\sqrt{\frac{\log p}{n}}\Big]R^2s^2c^{*4}\|\boldsymbol{b}_2-\boldsymbol{b}_2^*\|_2\|\boldsymbol{b}_1^+-\boldsymbol{b}_1^*\|_2.$$

Similarly, we can get the bound for T_2. This ends the proof. ■

3.3 Proof of Lemma 27

Recall that $\mathcal{P}^*$ is the dual norm of group lasso penalty $\mathcal{P}$. With a little abuse of notations, we define $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_n)^\top$ and $\mathcal{T}^*(\mathcal{X})=(\mathcal{T}^*(\mathcal{X}_1),\ldots,\mathcal{T}^*(\mathcal{X}_n))^\top$ in this section. According to the derivation of the gradient function in (12), we decompose the error by an spline approximation error term (T_1) and a statistical term (T_2) as follows,

$$\begin{aligned}
&\Big\|\nabla_1\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\Big\|_{\mathcal{P}^*} \\
= \;& \Big\|\frac{2}{n}\boldsymbol{F}^{1\top}(\boldsymbol{F}^1\boldsymbol{b}_1^* - \boldsymbol{y}) - \frac{2}{n}\boldsymbol{F}^{1\top}(\boldsymbol{F}^1\boldsymbol{b}_1^* - \boldsymbol{F}^{1*}\boldsymbol{b}_1^*)\Big\|_{\mathcal{P}^*} \\
= \;& \Big\|\frac{2}{n}\boldsymbol{F}^{1\top}(\boldsymbol{y} - \boldsymbol{F}^{1*}\boldsymbol{b}_1^*)\Big\|_{\mathcal{P}^*} = \Big\|\frac{2}{n}\boldsymbol{F}^{1\top}(\mathcal{T}^*(\mathcal{X}) - \boldsymbol{F}^{1*}\boldsymbol{b}_1^* + \boldsymbol{\epsilon})\Big\|_{\mathcal{P}^*} \\
\leq \;& \underbrace{\Big\|\frac{2}{n}\boldsymbol{F}^{1\top}(\mathcal{T}^*(\mathcal{X}) - \boldsymbol{F}^{1*}\boldsymbol{b}_1^*)\Big\|_{\mathcal{P}^*}}_{T_1} + \underbrace{\Big\|\frac{2}{n}\boldsymbol{F}^{1\top}\boldsymbol{\epsilon}\Big\|_{\mathcal{P}^*}}_{T_2}.
\end{aligned}$$

Step One: Bounding T_1. Denote $A_1 = \{j \in [p] | \|\boldsymbol{F}_j^1\|_2 \neq 0\}$. Since $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$, it's easy to see $|A_1| \leq C_0 s$ for some constant C_0 not depending on n, p, s. By the definition of dual norm $\mathcal{P}^*$ (See Definition 7), we obtain

$$\begin{aligned}
T_1 \;&=\; \Big\|\frac{2}{n}\sum_{i=1}^n \boldsymbol{F}_i^1\Big(\mathcal{T}^*(\mathcal{X}_i) - [\boldsymbol{F}^{1*}\boldsymbol{b}_1^*]_i\Big)\Big\|_{\mathcal{P}^*} \\
&=\; \max_{j\in A_1}\Big\|\frac{2}{n}\sum_{i=1}^n \boldsymbol{F}_{ij}^1\Big(\mathcal{T}^*(\mathcal{X}_i) - [\boldsymbol{F}^{1*}\boldsymbol{b}_1^*]_i\Big)\Big\|_2 \\
&\leq\; \max_{i\in[n]}\Big|\mathcal{T}^*(\mathcal{X}_i) - [\boldsymbol{F}^{1*}\boldsymbol{b}_1^*]_i\Big| \max_{j\in A_1}\Big\|\frac{2}{n}\sum_{i=1}^n \boldsymbol{F}_{ij}^1\Big\|_2. \qquad \text{(A7)}
\end{aligned}$$

Note that the first part of (A7) fully comes from the approximation error using B-spline basis functions for the nonparametric component. We bound T_1 in three steps as follows.

1. To bound the first part, we use Lemma 28 which quantifies the approximation error for a single component. To ses this, there exists a positive constant C_1 such that

$$\Big|f_{jkl}^{d_n}([\mathcal{X}_i]_{jkl}) - f_{jkl}^*([\mathcal{X}_i]_{jkl})\Big| \leq C_1 d_n^{-\kappa}, \;\; j, k, l \in [p].$$

   For the whole nonparametric function $\mathcal{T}^*$, we utilize the CP-low-rankness assumption (5) and group sparse assumption (7), which indicates

$$\max_{i\in[n]}\Big|\mathcal{T}^*(\mathcal{X}_i) - [\boldsymbol{F}^{1*}\boldsymbol{b}_1^*]_i\Big| = \max_{i\in[n]}\Big|\sum_{j=1}^p\sum_{k=1}^p\sum_{l=1}^p\Big(f_{jkl}^{d_n}([\mathcal{X}_i]_{jkl}) - f_{jkl}^*([\mathcal{X}_i]_{jkl})\Big)\Big| \leq C_1 s^3 d_n^{-\kappa}. \qquad \text{(A8)}$$

2. To bound the second part, by the definition of $\boldsymbol{F}_{ij}^1$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \boldsymbol{F}_{ij}^1 =& \Big(\frac{1}{n}\sum_{i=1}^n\langle\boldsymbol{\beta}_{211}\circ\boldsymbol{\beta}_{311}, [\mathcal{F}_1(\mathcal{X}_i)]_{j..}\rangle, \ldots, \frac{1}{n}\sum_{i=1}^n\langle\boldsymbol{\beta}_{2d_n1}\circ\boldsymbol{\beta}_{3d_n1}, [\mathcal{F}_{d_n}(\mathcal{X}_i)]_{j..}\rangle \\
&\ldots, \frac{1}{n}\sum_{i=1}^n\langle\boldsymbol{\beta}_{21R}\circ\boldsymbol{\beta}_{31R}, [\mathcal{F}_1(\mathcal{X}_i)]_{j..}\rangle, \ldots, \frac{1}{n}\sum_{i=1}^n\langle\boldsymbol{\beta}_{2d_nR}\circ\boldsymbol{\beta}_{3d_nR}, [\mathcal{F}_{d_n}(\mathcal{X}_i)]_{j..}\rangle\Big),
\end{aligned}$$

which implies that

$$\Big\|\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{F}_{ij}^{1}\Big\|_2 = \Big(\sum_{h=1}^{d_n}\sum_{r=1}^{R}\Big(\frac{2}{n}\sum_{i=1}^{n}\langle\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr},[\mathcal{F}_h(\mathcal{X}_i)]_{j\cdot\cdot}\rangle\Big)^2\Big)^{\frac{1}{2}}.$$

According to the property of B-spline basis function in Section 1, we have

$$\mathbb{E}\big\langle\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr},[\mathcal{F}_h(\mathcal{X}_i)]_{j\cdot\cdot}\big\rangle \le d_n^{-1}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}\beta_{3hrl} \le \frac{C_0 s}{d_n}\Big(\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}, \tag{A9}$$

where the second inequality comes from Cathy-Schwarz inequality and sparsity assumption on $\boldsymbol{b}_2, \boldsymbol{b}_3$. On the other hand, recall that $\sup_x|\psi_{jklh}(x)| \le 1$ for all $j,k,l \in [p]$. With the rotation invariance, the ϕ_2-Orlicz norm of $\langle\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr},[\mathcal{F}_h(\mathcal{X}_i)]_{j\cdot\cdot}\rangle$ can be bounded by $(\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}^2\beta_{3hrl}^2)^{\frac{1}{2}}$. Combining (A9) and Hoeffding-type concentration inequality (See Lemma 35), we have with probability at least $1-1/n$,

$$\frac{2}{n}\sum_{i=1}^{n}\langle\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr},[\mathcal{F}_h(\mathcal{X}_i)]_{j\cdot\cdot}\rangle \le 2\Big(\frac{C_0 s}{d_n}+\sqrt{\frac{\log(en)}{n}}\Big)\Big(\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}, \tag{A10}$$

which implies

$$\max_{j\in A_1}\Big\|\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{F}_{ij}^{1}\Big\|_2 \le 2\Big(\frac{C_0 s}{d_n}+\sqrt{\frac{\log(en)}{n}}\Big)\Big(\sum_{h=1}^{d_n}\sum_{r=1}^{R}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}, \tag{A11}$$

with probability at least $1-Rd_n s/n$.

3. Putting (A8)-(A11) together, we obtain

$$T_1 \le 2C_1 s^3 d_n^{-\kappa}\Big(\frac{C_0 s}{d_n}+\sqrt{\frac{\log(en)}{n}}\Big)\Big(\sum_{h=1}^{d_n}\sum_{r=1}^{R}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta_{2hrk}^2\beta_{3hrl}^2\Big)^{\frac{1}{2}}. \tag{A12}$$

with probability at least $1-Rd_n s/n$ for some absolute constant C_0, C_1.

Step Two: Bounding T_2. Recall that $\boldsymbol{F}^{1\top}\boldsymbol{\epsilon} = (\boldsymbol{F}_1^{1\top}\boldsymbol{\epsilon},\ldots,\boldsymbol{F}_p^{1\top}\boldsymbol{\epsilon})^\top \in \mathbb{R}^{pd_n\times 1}$. Then,

$$\begin{aligned}
T_2 &= \max_{j\in A_1}\Big\|\frac{2}{n}\boldsymbol{F}_j^{1}\boldsymbol{\epsilon}\Big\|_2 = \max_{j\in A_1}\Big\|\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{F}_{ij}^{1}\epsilon_i\Big\|_2\\
&\le \frac{1}{\sqrt{n}}\max_{j\in A_1,h\in[d_n],r\in[R]}\sqrt{\frac{d_n}{n}}\sum_{i=1}^{n}\epsilon_i\langle\boldsymbol{\beta}_{2hr}\circ\boldsymbol{\beta}_{3hr},[\mathcal{F}_h(\mathcal{X}_i)]_{j\cdot\cdot}\rangle\\
&= \frac{1}{\sqrt{n}}\max_{j\in A_1,h\in[d_n],r\in[R]}\sum_{k\in w(\boldsymbol{b}_2)}\sum_{l\in w(\boldsymbol{b}_3)}\sqrt{\frac{d_n}{n}}\sum_{i=1}^{n}\epsilon_i\psi_{jklh}(\mathcal{X}_i)\beta_{2hrk}\beta_{2hrl}.
\end{aligned}$$

where the definition of $w(\boldsymbol{x})$ is presented in the beginning of Section 4.1. From initial value assumption and Condition 17, we have

$$|\beta_{2hrk}-\beta_{2jrk}^*| \;\le\; \max_{h,k}|\beta_{2hrk}-\beta_{2hrk}^*| \le \|\beta_{2hrk}-\beta_{2hrk}^*\|_2 \le \alpha,$$

and thus $\beta_{2hrk} \le \beta^*_{2hrk} + c^*$. The same result holds for β_{3hk}. Therefore, by applying Lemma 36, we have

$$\begin{aligned} T_2 &\le \frac{s^2}{\sqrt{n}}(\alpha + c^*)^2 \max_{j\in A_1, h\in[d_n], r\in[R]} \sqrt{\frac{d_n}{n}} \sum_{i=1}^n \epsilon_i \psi_{jklh}(\mathcal{X}_i) \\ &\le C_3 \sigma \frac{s^2\sqrt{\log(pd_n)}}{\sqrt{n}}, \end{aligned} \tag{A13}$$

with probability at least $1 - 4C_0Rs/n$, where σ is the noise level.

Step Three: Summary. Putting the bounds (A12) and (A13) together, we obtain that with probability at least $1 - C_0Rd_ns/n$,

$$\begin{aligned} &\left\|\nabla_1\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\right\|_{\mathcal{P}^*} \\ \le\ & \Big[C_1 s^3 d_n^{-\kappa}\Big(\frac{C_0 s}{d_n} + \sqrt{\frac{\log ep}{n}}\Big)\Big]\Big(\sum_{h=1}^{d_n}\sum_{r=1}^{R}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta^2_{2hrk}\beta^2_{3hrl}\Big)^{\frac{1}{2}} + C_3\sigma\frac{s^2\sqrt{\log(pd_n)}}{\sqrt{n}} \\ \le\ & \Big[\frac{C_1 s^3}{d_n^{\kappa}}\sqrt{\frac{\log ep}{n}} + \frac{C_2 s^4}{d_n^{\kappa+1}}\Big]\Big(\sum_{h=1}^{d_n}\sum_{r=1}^{R}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta^2_{2hrk}\beta^2_{3hrl}\Big)^{\frac{1}{2}} + C_3\sigma\frac{s^2\sqrt{\log(pd_n)}}{\sqrt{n}}, \end{aligned}$$

where C_1, C_2, C_3 are some positive constants. According to Condition 17 and $\boldsymbol{b}_2 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_2^*)$, $\boldsymbol{b}_3 \in \mathcal{B}_{\alpha,s}(\boldsymbol{b}_3^*)$,

$$\sum_{h=1}^{d_n}\sum_{r=1}^{R}\sum_{k=1}^{p}\sum_{l=1}^{p}\beta^2_{2hrk}\beta^2_{3hrl} = \sum_{h=1}^{d_n}\Big(\sum_{k=1}^{p}\beta^2_{jhk}\Big)\Big(\sum_{l=1}^{p}\beta^2_{3hl}\Big) \le Rd_n^{1/2}s^2c^{*4}.$$

By setting $C_1 = \max\{C_1, C_2, C_2\}$, we have with probability at least $1 - C_0Rd_ns/n$,

$$\begin{aligned} &\left\|\nabla_1\mathcal{L}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3) - \nabla_1\widetilde{\mathcal{L}}(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \boldsymbol{b}_3)\right\|_{\mathcal{P}^*} \\ \le\ & C_1Rc^{*4}\Big[\frac{s^5}{d_n^{\kappa-1/2}}\sqrt{\frac{\log(en)}{n}} + \frac{s^6}{d_n^{\kappa+1/2}} + \sigma\sqrt{\frac{s^4\log(pd_n)}{n}}\Big]. \end{aligned}$$

This ends the proof. ■

4. Supporting Lemmas

Lemma 35 (Hoeffding-type inequality) *Suppose $\{X_i\}_{i=1}^n$ are i.i.d sub-Gaussian random variable with $\|X_i\|_{\phi_2} \le K$, where K is an absolute constant. For fixed $\boldsymbol{a} \in \mathbb{R}^n$, we have w.p.a $1 - \delta$,*

$$\Big|\sum_{i=1}^n a_iX_i - \mathbb{E}(\sum_{i=1}^n a_iX_i)\Big| \le C_0K\|\boldsymbol{a}\|_2\sqrt{\log(e/\delta)}.$$

Lemma 36 (Lemma 2 in Huang et al. (2010)) *Suppose that Condition 15-16 hold. Let*

$$T_{jkl} = \sqrt{\frac{d_n}{n}} \sum_{i=1}^{n} \psi_{jklh}([\mathcal{X}_i]_{jkl})\epsilon_i, \text{ for } j \in [p], k \in [p], l \in [p], h \in [d_n],$$

and $T_n = \max_{j,k,l\in[p],h\in[d_n]} |T_{jkl}|$*. When* $d_n\sqrt{pd_n}/n \to 0$*, we have for some constant* C_1*,*

$$\mathbb{E}(T_n) = C_1\sqrt{\log(pd_n)}.$$