Abstract

We propose a novel broadcasting idea to model the nonlinearity in tensor regression non-parametrically. Unlike existing non-parametric tensor regression models, the resulting model strikes a good balance between flexibility and interpretability. A penalized estimation and corresponding algorithm are proposed. Our theoretical investigation, which allows the dimensions of the tensor covariate to diverge, indicates that the proposed estimation enjoys a desirable convergence rate. We also provide a minimax lower bound, which characterizes the optimality of the proposed estimator in a wide range of scenarios. Numerical experiments are conducted to confirm the theoretical finding and show that the proposed model has advantages over existing linear counterparts.

Keywords: Nonlinear regression; Tensor low rank; Polynomial splines; Region selection; Elastic-net penalization.

1 Introduction

Recent years have witnessed a massive emergence of tensor data in many different areas, such as clinical applications (Wang et al., 2014), computer vision (Lu et al., 2013), genomics (Durham et al., 2018), neuroscience (Zhou et al., 2013), and recommender systems (Zhu et al., 2018). Uncovering relationships among different variables from tensor data often lead to enhanced understanding of scientific and engineering problems. One recent statistical development under this setup is tensor regression (Zhou et al., 2013). In this work, we focus on models that involve a tensor covariate $X = (X_{i_1,i_2,...,i_D}) \in \mathbb{R}^{p_1 \times p_2 \times ... \times p_D}$ of order $D$. In passing, it is also worth mentioning that regression of a tensor response on a vector covariate (e.g., Sun and Li, 2017; Li and Zhang, 2017; Hu et al., 2019) is also a popular research direction.

Commonly seen are three major types of tensor regression, with different forms of response. The first is scale-on-tensor regression, i.e., the response is a scalar (Zhou et al., 2013; Zhao et al., 2014; Hou et al., 2015; Chen et al., 2019). Within this category, there are methods that focus particularly on image covariates (Reiss and Ogden, 2010; Zhou and Li, 2014; Wang et al., 2017; Kang et al., 2018). The second is vector-on-tensor regression, in which we have a vector response (Miranda et al., 2018). The last one is tensor-on-tensor regression with a tensor output (Hoff, 2015; Lock, 2018; Raskutti et al., 2019).

Most of the aforementioned models make a strong assumption that the tensor covariate is able to predict the response through (known transformations of) linear functions. To
date, very few work go beyond this limitation. On the application side, Zhao et al. (2014) and Hou et al. (2015) used Gaussian processes to model potential nonlinear effects of tensor covariates in video surveillance applications and neuroimaging analyses. Their methods are geared for prediction, but lack for interpretability and theoretical justification. Moreover, the performance of this approach heavily relies upon the choice of kernel function, which is not easy to design for efficiently harnessing the tensor structure.

Another class of methods incorporates nonlinearity through a more explicit function space by imposing low-rank structures on covariates. Kanagawa et al. (2016) considered a regression model with respect to a rank-one tensor covariate, i.e., \( X = x_1 \odot x_2 \odot \cdots \odot x_D \), where \( \odot \) denotes the outer product and \( x_d \) is a \( p_d \)-dimensional vector, \( d = 1, \ldots, D \). Imaizumi and Hayashi (2016) extended this work to a higher-rank tensor and proposed the model

\[
m(X) = \sum_{r=1}^{R} \sum_{q=1}^{Q} \lambda_q \prod_{d=1}^{D} g_{d,r}(x_{q,d}),
\]

where \( X \) is assumed to have a smallest CANDECOMP/PARAFAC (CP) decomposition

\[
X = \sum_{q=1}^{Q} \lambda_q x_{q,1} \odot x_{q,2} \odot \cdots \odot x_{q,D},
\]

where the Euclidean norm \( \|x_{q,d}\|_2 = 1 \) and \( \lambda_Q \geq \lambda_{Q-1} \geq \cdots \geq \lambda_1 \geq 0 \). When \( Q = 1 \), (1) recovers the model of Kanagawa et al. (2016). In order to significantly reduce the number of unknown functions to be estimated, a small value of \( Q \) is usually recommended. However, in most cases, the tensor covariate is not exactly low-rank, and the rank of the covariate varies from observation to observation within the data set. Furthermore, although the additive form of (1) has reduced the model complexity, simultaneously estimating \( DR \) unknown multivariate functions, \( g_{d,r}'s \), remains a challenging problem. For example, given a \( 64 \times 64 \times 64 \) 3D-image covariate \((p_1 = p_2 = p_3 = 64)\), we need to estimate \( 3R \) unknown 64-dimensional functions, which will lead to the curse of dimensionality. This aligns with a finding, from Imaizumi and Hayashi (2016), that the asymptotic convergence rate of this model grows exponentially with \( \max_d p_d \). Finally, this model is difficult to interpret since the nonlinear modeling is directly built upon the CP representation of the covariates, which may not be unique (Stegeman and Sidiropoulos, 2007). As pointed out by a reviewer, a recent manuscript (Hao et al., 2019) proposed a sparse additive tensor regression model, which can be regarded as a generalization of the spline approximation (5) of our model (3). Like the aforementioned models, the added flexibility would likely result in better predictions. However, their model is based upon a low-rank assumption on the finite-dimensional tensor of spline coefficients. As the number of spline basis functions changes, the interpretation of the low-rank structure also varies, which results in ambiguity in the target nonparametric function class and a difficult interpretation.

Therefore, although these existing nonlinear models demonstrate successes in certain applications, they suffer from the curse of dimensionality and/or possess weak interpretability. In this article, we propose an alternative that addresses both of these issues. Our proposed model extends the low-rank tensor linear model developed by Zhou et al. (2013), which we briefly describe as follows. Given a vector covariate \( z \in \mathbb{R}^{p_0} \), a tensor covariate \( X \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D} \), and a response variable \( y \in \mathcal{Y} \subseteq \mathbb{R} \), Zhou et al. (2013) proposed a generalized tensor linear
model through a predetermined link function \( g \)

\[
g\{E(y|z, X)\} = \nu + \gamma^T z + \langle B, X \rangle,
\]

where \( \nu \in \mathbb{R}, \gamma \in \mathbb{R}^{p_0}, B \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D} \) are unknown parameters, and \( \langle \cdot, \cdot \rangle \) denotes the componentwise inner product, i.e., \( \langle B, X \rangle = \sum_{i_1,\ldots,i_D} B_{i_1,\ldots,i_D} X_{i_1,\ldots,i_D} \). In particular, the coefficient tensor \( B \) is assumed to admit a CP decomposition

\[
B = \sum_{r=1}^R \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D},
\]

where \( \beta_{r,d} \in \mathbb{R}^{p_d} \) and \( R \) is the CP rank. Combined with sparsity-inducing regularization, Zhou et al. (2013) and Zhou and Li (2014) showed that low-rank coefficient tensor \( B \) can be used to identify the regions (entries) of \( X \) that are relevant to predict the response variable. Apart from CP low-rank structures, we note that another popular and successful low-rank modeling is based on multilinear rank (e.g., Li et al., 2018; Zhang et al., 2020). In this paper, we will only focus on the low-rank modeling via CP rank.

In many real-world applications, entries within some regions of the tensor (especially images) share similar effects due to certain spatial structures. For examples, Zhou et al. (2013) and Miranda et al. (2018) both provided evidences that brains demonstrate spatially clustered effects on certain diseases. Motivated by these observations and the possibility of nonlinear effects, we propose to “broadcast” similar nonlinear relationships to different entries of the tensor covariate. On a high-level, we model the nonlinear effects by uni-dimensional nonparametric functions, which are supposed to be applied to an individual entry. These uni-dimensional functions are then shared by every entry to indicate the clustered effect. We call this operation of distributing a uni-dimensional function to all entries “broadcasting”. Additional scaling coefficients are used to linearly scale the effects of the uni-dimensional functions. Through regularizing these scaling coefficients, we are able to restrict the effects of certain uni-dimensional functions to smaller regions. As shown by Zhou et al. (2013) and Zhou and Li (2014), lasso-type regularization alone may result in poor performance in region selection, while an additional low-rank constraint/regularization produce more successful results. Therefore we also restrict the scaling coefficient to be low-rank. Combined with multiple broadcasting, the proposed model can produce reasonably complex and interpretable structures, as shown in Section 2.2.

Within the proposed model, all aforementioned ideas are integrated into a (penalized) least squares framework. We develop an alternative updating algorithm as well as the asymptotic rates of convergence for the proposed estimations. Our theory includes tensor linear model (Zhou et al., 2013) as a special case. Unlike Zhou et al. (2013), ours is of high-dimensional nature, which allows \( p_1, \ldots, p_D \) to diverge with the sample size. We believe this asymptotic framework is more relevant to many applications, where the data, such as images, involve large values of \( p_j \)’s as compared to the sample size. To construct the asymptotic analysis, we have provided a novel restricted eigenvalue result, as well as a new entropy bound. To characterize the optimality, we have also obtained a minimax lower bound, whose rate matches with that of the error upper bound of our estimator in a wide range of scenarios. Through a real data example, we demonstrate the power of the proposed broadcasted nonparametric tensor regression. Overall, the proposed method timely responds to a number of growing needs of
modeling nonlinearity with interpretable models and rigorous theoretical developments for tensor data.

The rest of this paper is organized as follows. Section 2 introduces the broadcasted non-parametric model. The proposed estimation method with the algorithm and the corresponding theoretical results are respectively presented in Sections 3 and 4. The practical performance of the proposed method illustrated via both a simulation study and a real data application can be found in Section 5. The main contributions of this paper are summarized in Section 6 with some concluding remarks. Technical details are provided in a separate online supplemental document.

2 Model

Consider a tensor covariate \( X \in \mathcal{X} := \{(A_{i_1,\ldots,i_D})_{i_1,\ldots,i_D=1}^{p_1,\ldots,p_D} : A_{i_1,\ldots,i_D} \in I\} \), where \( I \) is a compact subset of \( \mathbb{R} \). Unless otherwise specified, we assume \( I = [0,1] \) without loss of generality. Throughout this paper, we focus on the general model

\[
y = m(X) + \epsilon,
\]

where \( m : \mathcal{X} \to \mathbb{R} \) is an unknown regression function of interest and \( \epsilon \) is a random error with mean zero. The observed data \( \{(y_i, X_i)\}_{i=1}^n \) are modeled as i.i.d. copies of \((y, X)\). In this section, we propose an interpretable model for the regression function \( m(\cdot) \).

2.1 Common nonparametric strategies: curse of dimensionality

As discussed in Section 1, existing work of nonparametric tensor regression is not only lacking in interpretability but also suffering from a slow rate of convergence due to the curse of dimensionality. This issue of dimensionality also occurs if one adopts other common nonparametric regression models, such as additive models, by directly flattening the tensor covariate into a vector. To relate with the standard nonparametric regression methods, we briefly discuss them here, and highlight the issue of dimensionality, which motivates the proposed model in Section 2.2.

One of the most general models for the regression function \( m(\cdot) \) is an unstructured (smooth) mapping from \( \mathcal{X} \) to \( \mathbb{R} \). Despite its flexibility, this model suffers heavily from high dimensionality. For a typical \( 64 \times 64 \times 64 \) image, we are facing a nonparametric estimation of a function with dimension \( 64^3 \), which is generally impractical.

A common alternative in the literature of nonparametric regression is an additive form of the regression function (Stone, 1985; Hastie and Tibshirani, 1990; Wood, 2017), i.e.,

\[
m(X) = \sum_{i_1, i_2, \ldots, i_D} m_{i_1 i_2 \ldots i_D}(X_{i_1 i_2 \ldots i_D}),
\]

where \( m_{i_1 i_2 \ldots i_D} \)'s are unknown uni-dimensional functions. This model however needs to simultaneously estimate \( s = \prod_{d=1}^D p_d \) uni-dimensional functions, in which consistent estimation is generally impossible for \( s \geq n \). In this case, the sparsity assumption (Lin and Zhang, 2006; Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010; Raskutti et al., 2012; Fan et al., 2011; Chen et al., 2018) could help obtain consistent estimation of the regression function. Nevertheless, general sparse estimators, when applied to a vectorized tensor covariate, would
ignore the potential tensor structure and might result in large bias, especially when the sample size \( n \) is much smaller than \( s \).

Another common modeling is the single index model (Ichimura, 1993; Horowitz and Härdle, 1996), in the sense that

\[
m(X) = g \left( \sum_{i_1, \ldots, i_D} a_{i_1,i_2,\ldots,i_D} X_{i_1,i_2,\ldots,i_D} \right),
\]

where \( g \) is an unknown uni-dimensional function and \( a_{i_1,i_2,\ldots,i_D} \)'s are \( s \) unknown weight parameters. Although there is only one uni-dimensional function to be estimated, this model involves abundant number of weight parameters, sometimes larger than the sample size. One could also impose sparsity assumption on the weight parameters. The readers are referred to Alquier and Biau (2013), Radchenko (2015), and references therein for more details on this approach. However, similar issue of ignoring tensor structures would also show up. Such problem would be aggravated in more complicated index models such as the additive index model and multiple indices model.

We propose a novel and economical model which makes use of the tensor structure. Our model is closely related to the additive models, but has overcome the aforementioned problems.

### 2.2 Low-rank modeling with broadcasting

As mentioned above, the additive models involve too many functions. A simple remedy is to restrict all entries to share the same function:

\[
m(X) = \frac{1}{s} \sum_{i_1,i_2,\ldots,i_D} f(X_{i_1,i_2,\ldots,i_D}),
\]

where \( f \) is a uni-dimensional function residing in a function class \( \mathcal{H} \) to be specified later, and the scaling \( s^{-1} \) is introduced to match with our proposed model (3). In other words, we broadcast\(^1\) the same function \( f \) to every entry. We formally define the broadcasting operator \( \mathcal{B} : \mathcal{H} \times \mathcal{X} \rightarrow \mathbb{R}^{p_1 \times \cdots \times p_D} \) by

\[
(\mathcal{B}(f,X))_{i_1,i_2,\ldots,i_D} = f(X_{i_1,i_2,\ldots,i_D}), \quad \text{for all } i_1, \ldots, i_D.
\]

Figure 1 depicts an example of the broadcasting operation. In many real life applications, entries within some regions of the tensor (especially images) share similar effects due to certain spatial structures such as a spatially clustered effect. For instance, Zhou et al. (2013) showed that voxels within two brain subregions have similar linkages with attention deficit hyperactivity disorder. Miranda et al. (2018) demonstrated that voxels within several subregions of the brain have a spatially clustered effect on Alzheimer’s disease. Hence, broadcasting a nonlinear relationship (with the response) is a well-motivated modeling strategy. But the assumption that every entry has the same nonlinear effect on the response is very restrictive. Specifically, in many image data, there are usually only one or a few clusters of entries that are related to the response. Therefore, we move beyond a simple broadcasting structure to achieve more adaptive modeling.

\(^1\)A term widely used for similar operations in programming languages such as Python.
An example of the Figure 2: in scaling tensor. The white elements in effects adapted to different subregions. We give two simple examples of effect in Figure 2. Several components can be combined to characterize different nonlinear be made specifically concentrated on a subregion of the tensor. We demonstrate the scaling model is unidentifiable, i.e., the broadcasted functions and scaling tensors are not uniquely shared nonlinear effect (functions; in the right panel, there is a rank-two region formed by two scaling tensors. In the left panel, there are two rank-one regions (shaded) with different nonlinear Figure 3, where the shaded regions correspond to nonzero entries in the corresponding scaling tensors are of rank different components, each of which is composed of a uni-dimensional function \( f_r \) to be broadcasted, and a rank-one scaling (coefficient) tensor \( \beta_{r,1} \circ \cdots \circ \beta_{r,D} \) to linearly scale the effect across different entries. The model is economical since these broadcasted functions are uni-dimensional and these scaling tensors are of rank 1.

If an appropriate sparse estimation is imposed on the scaling tensors, a component can be made specifically concentrated on a subregion of the tensor. We demonstrate the scaling effect in Figure 2. Several components can be combined to characterize different nonlinear effects adapted to different subregions. We give two simple examples of \( D = 2 \) depicted in Figure 3, where the shaded regions correspond to nonzero entries in the corresponding scaling tensors. In the left panel, there are two rank-one regions (shaded) with different nonlinear functions; in the right panel, there is a rank-two region formed by two scaling tensors with a shared nonlinear effect (\( f_1 = f_2 \)).

Similar to the tensor linear model (Zhou et al., 2013), the parameterization in the proposed model is unidentifiable, i.e., the broadcasted functions and scaling tensors are not uniquely determined. For instance, one can multiply \( \beta_{r,1} \) by 10, and divide \( \beta_{r,2} \) by 10, while still
obtain the same \( m(\cdot) \). Another example is a permutation of the components. However, to understand the nonlinear effect of entries, only the identification of \( m(\cdot) \) is needed and thus such non-identifiability is in general not an issue. In particular, we are able to directly study the asymptotic behaviors of the estimations of \( m(\cdot) \) in Section 4. For computation, on the other hand, some of these identifiability issues lead to algorithmic instability and so several restrictions are introduced in Section 3 to obtain an efficient algorithm. For a complete discussion on parameter identification, we refer interested readers to Section C of the supplementary material (SM), where sufficient conditions similar to Kruskal’s uniqueness condition (Kruskal, 1989) are provided.

3 The proposed estimator and its computation

3.1 Spline approximation and penalized estimation

The broadcasted functions \( f_r, r = 1, \ldots, R \), will be approximated by B-spline functions of order \( \zeta \), i.e.,

\[
f_r(x) \approx \sum_{k=1}^{K} \alpha_{r,k} b_k(x),
\]

where \( b(x) = (b_1(x), \ldots, b_K(x))^T \) is a vector of B-spline basis functions and \( \alpha_{r,k} \)'s are the corresponding spline coefficients. By writing \( \alpha_r = (\alpha_{r,1}, \ldots, \alpha_{r,K})^T \) and ignoring the spline approximation error, the regression function (3) can be approximated by

\[
m(X) \approx \nu + \frac{1}{s} \sum_{r=1}^{R} \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \alpha_r, \Phi(X) \rangle,
\]

where \( \Phi: \mathcal{X} \to \mathbb{R}^{p_1 \times \cdots \times p_D \times K} \) is defined by \( (\Phi(X))_{i_1, \ldots, i_D, k} = b_k(X_{i_1, \ldots, i_D}) \). We remark that (5) is not the proposed model, but merely an approximation of the proposed nonparametric model (3). To separate out the constant effect from \( f_r \)'s, we impose the conditions \( \int_0^1 f_r(x)dx = 0 \), \( r = 1, \ldots, R \), which lead to

\[
\int_0^1 \sum_{k=1}^{K} \alpha_{r,k} b_k(x)dx = 0, \quad r = 1, \ldots, R.
\]
Let \( u_k = \int_0^1 b_k(x) dx \). We consider the following optimization problem

\[
\arg \min_{\nu, \mathbf{A}} \sum_{i=1}^n \left( y_i - \nu - \frac{1}{s} \langle \mathbf{A}, \Phi(X_i) \rangle \right)^2
\]

\[
\text{s.t. } \mathbf{A} = \sum_{r=1}^R \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \alpha_r
\]

\[
\sum_{k=1}^K \alpha_{r,k} u_k = 0, \quad r = 1, \ldots, R,
\]

and the estimated regression function as

\[
\hat{m}_{LS}(X) = \hat{\nu}_{LS} + \frac{1}{s} \langle \hat{A}_{LS}, \Phi(X) \rangle,
\]

where \((\hat{A}_{LS}, \hat{\nu}_{LS})\) is a solution of (7).

Directly solving (7) is not computationally efficient since it involves too many linear constraints. To further simplify the optimization problem, we remove the constraints by using an equivalent truncated power basis (Ruppert et al., 2003). We let \( \{\tilde{b}_k(x)\}_{k=1}^K \) denote the truncated power basis:

\[
\tilde{b}_1(x) = 1, \quad \tilde{b}_2(x) = x, \ldots, \tilde{b}_\zeta(x) = x^{\zeta-1},
\]

\[
\tilde{b}_{\zeta+1}(x) = (x - \xi_2)^{\zeta-1}, \ldots, \tilde{b}_K(x) = (x - \xi_{K-\zeta+1})^{\zeta-1},
\]

where \( \zeta \) and \((\xi_2, \ldots, \xi_{K-\zeta+1})\) are the order and the interior knots of the aforementioned B-spline. Using these basis functions, we consider the optimization

\[
\arg \min_{\tilde{\nu}, \tilde{\mathbf{A}}} \sum_{i=1}^n \left( y_i - \tilde{\nu} - \frac{1}{s} \langle \tilde{\mathbf{A}}, \tilde{\Phi}(X_i) \rangle \right)^2
\]

\[
\text{s.t. } \tilde{\mathbf{A}} = \sum_{r=1}^R \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \tilde{\alpha}_r,
\]

where \( \tilde{\Phi} : \mathcal{X} \rightarrow \mathbb{R}^{p_1 \times \cdots \times p_D \times (K-1)} \) is defined by \( \langle \tilde{\Phi}(X) \rangle_{i_1, \ldots, i_D, k} = \tilde{b}_{k+1}(X_{i_1, \ldots, i_D}) \), \( k = 1, \ldots, K-1 \), and \( \tilde{\alpha}_r \in \mathbb{R}^{K-1} \) is the vector of coefficients. Compared with (7), the mean zero constraints are removed by reducing one degree of freedom in the basis functions. Lemma B.1 in Section B.1 of the SM shows that the optimization (8) results in the same estimated regression function, i.e.,

\[
\hat{m}_{LS}(X) = \hat{\nu}_{LS} + \frac{1}{s} \langle \hat{\mathbf{A}}_{LS}, \tilde{\Phi}(X) \rangle,
\]

where \((\hat{\nu}_{LS}, \hat{\mathbf{A}}_{LS})\) is a solution of (8).

To improve estimation performance (when sample size is relatively small) as well as to enhance interpretability, we add an additional penalty term to the optimization. In particular,
a penalized estimation is proposed by solving

$$\arg \min_{\tilde{\nu}, \tilde{A}} \sum_{i=1}^{n} \left( y_i - \tilde{\nu} - \frac{1}{s} \langle \tilde{A}, \tilde{\Phi}(X_i) \rangle \right)^2 + G(\theta)$$

s.t. $$\tilde{A} = \sum_{r=1}^{R} \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \tilde{\alpha}_r$$

$$\|\tilde{\alpha}_r\|_2^2 = 1, \quad r = 1, \ldots, R,$$

where $$G(\theta)$$ is a penalty function (that may depend on the sample size $$n$$) and

$$\theta = (B_1, \ldots, B_D), \quad B_d = (\beta_{1,d}, \ldots, \beta_{R,d}), \quad d = 1, \ldots, D.$$

Corresponding estimated regression function $$\hat{m}_{\text{PLS}}$$ can be reconstructed similarly as (9). Obviously, $$\hat{m}_{\text{PLS}} = \hat{m}_{\text{LS}}$$ if $$G(\theta) \equiv 0$$. For simplicity, we focus on penalty function in an additive form:

$$G(\theta) = \sum_{d=1}^{D} P_d(\beta_{1,d}, \ldots, \beta_{R,d}),$$

where $$P_d(\cdot)$$ is some penalty function. The technical requirement of the penalty is detailed in Assumption 1 below. Typical choices of the penalty function $$P_d(\cdot)$$ include the lasso penalty (Tibshirani, 1996), the elastic-net penalty (Zou and Hastie, 2005), and the group lasso penalty (Yuan and Lin, 2006). Note that the magnitudes of $$\beta_{r,1}, \ldots, \beta_{r,D}$$ and $$\tilde{\alpha}_r$$ are not identified. Penalization on $$\beta_{r,d}'s$$ would enlarge $$\tilde{\alpha}_r$$. The unit-norm restrictions for $$\tilde{\alpha}_r's$$ are introduced to prevent these scaling issues.

### 3.2 Algorithm

We propose a scale-adjusted block-wise descent algorithm to solve (10). Recall $$B_d = (\beta_{1,d}, \ldots, \beta_{R,d})$$, $$d = 1, \ldots, D$$. Analogously, we denote $$\hat{B}_{D+1} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_R)$$. For convenience, we write the squared loss and the whole objective function as

$$L(\tilde{\nu}, \theta, \hat{B}_{D+1}) = \sum_{i=1}^{n} \left( y_i - \tilde{\nu} - \frac{1}{s} \langle \tilde{A}, \tilde{\Phi}(X_i) \rangle \right)^2,$$

and $$LG(\hat{\nu}, \hat{\theta}, \hat{B}_{D+1}) = L(\hat{\nu}, \theta, \hat{B}_{D+1}) + G(\theta)$$, respectively. Before proposing the algorithm, we first rearrange $$\langle \cdot, \cdot \rangle$$ in $$LG$$ using the Khatri-Rao product and mode-$$d$$ matricization. The Khatri-Rao product is defined as a column-wise Kronecker product for two matrices with the same column number (Smilde et al., 2005). More precisely, let $$B = (b_1, \ldots, b_L) \in \mathbb{R}^{I \times L}$$ and $$B' = (b'_1, \ldots, b'_L) \in \mathbb{R}^{J \times L}$$ be two generic matrices which have the same number of columns, their Khatri-Rao product $$B \odot B' \in \mathbb{R}^{IJ \times L}$$ is defined as

$$B \odot B' = [b_1 \otimes b'_1 \quad b_2 \otimes b'_2 \quad \cdots \quad b_L \otimes b'_L],$$

where $$\otimes$$ denotes the Kronecker product (Kolda and Bader, 2009). The mode-$$d$$ matricization of a tensor $$A \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_D}$$, denoted by $$A_{(d)}$$, maps a tensor into a matrix according to its
We can multiply alternatively update $\tilde{\Phi}(\alpha_r,\hat{\Phi}(X))$ and Bader, 2009) and are able to obtain the closed form solutions equivalent parametrization. For case, as described in Section B.2 of the SM, (13) can be written as a convex problem in an $\lambda$ where

$$\sum_{r=1}^{R} \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \hat{\alpha}_r, \tilde{\Phi}(X) \rangle = \langle B_d, \hat{\Phi}(X)(d)B_{-d} \rangle = \langle \text{vec}(\hat{\Phi}(X)(d)B_{-d}), \text{vec}(B_d) \rangle,$$

where $B_{-d} = B_1 \odot \cdots \odot B_{d-1} \odot B_{d+1} \odot \cdots \odot \tilde{B}_{D+1}$, vec($\cdot$) is a vectorization operator (Kolda and Bader, 2009) and $\hat{\Phi}(X)(d)$ is the mode-$d$ matricization of tensor $\hat{\Phi}(X)$. We can thus alternatively update $B_d$, $d = 1, \cdots, D$, by a penalized linear regression. As for $\tilde{B}_{D+1}$, it can be updated by optimization method on the oblique manifold (Selvan et al., 2012) due to the restriction that the norm of each column of $\tilde{B}_{D+1}$ is 1.

Of special attention is the magnitude shift among $\beta_{r,d}$’s for $d = 1, \cdots, D$. As an example, we can multiply $\beta_{r,d_1}$ by 10 and divide $\beta_{r,d_2}$ by 10, $d_1 \neq d_2$, without changing the value of the squared loss. This manipulation can, however, change the value of the penalty $G(\theta)$. To improve the algorithmic convergence, we propose a rescaling strategy for the penalty in the form of (12). For $r = 1, \cdots, R$, we solve the following optimization problem

$$\arg \min_{\rho_{r,d} : r = 1, \ldots, R, d = 1, \ldots, D} \sum_{d=1}^{D} P_d(\rho_{1,d}\beta_{1,d}, \ldots, \rho_{R,d}\beta_{R,d})$$

s.t. $\prod_{d=1}^{D} \rho_{r,d} = 1$ and $\rho_{r,d} > 0$,

$$\rho_{r,d} = 1$$ if $\beta_{r,d} = 0$, for $r = 1, \ldots, R, d = 1, \ldots, D,$

and we replace $\beta_{r,d}$ by $\hat{\rho}_{r,d}\beta_{r,d}$ at the end of each iterative step of solving (10) (see Algorithm 1), where $\{\hat{\rho}_{r,d} : r = 1, \ldots, R, d = 1, \ldots, D\}$ is the minimizer of (13). This replacement step never increases the objective value (as shown in Proposition 1 below). Our theory and algorithm work under a penalty in the additive form (12) and fulfilling Assumption 1. To avoid complication, we defer the discussion of the general cases to Section B.2 in the SM.

Now, we provide more discussion of the elastic-net penalty

$$P_d(\beta_{1,d}, \ldots, \beta_{R,d}) = \lambda_1 \sum_{r=1}^{R} \left\{ \frac{1}{2} (1 - \lambda_2) \|\beta_{r,d}\|_2^2 + \lambda_2 \|\beta_{r,d}\|_1 \right\},$$

where $\lambda = (\lambda_1, \lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 \in [0, 1]$, and $\lambda$ is allowed to depend on $n$. In this case, as described in Section B.2 of the SM, (13) can be written as a convex problem in an equivalent parametrization. For $\lambda_2 \in (0, 1)$, the method of Lagrange multipliers and Newton’s method can be used to solve (13). While for the special boundary cases, i.e., $\lambda_2 \in \{0, 1\}$, we are able to obtain the closed form solutions

$$\hat{\rho}_{r,d} = \begin{cases} 
\frac{1}{\|\beta_{r,d}\|_1} \prod_{d=1}^{D} \|\beta_{r,d}\|_{1/D}^{1/D}, & \text{if } \lambda_2 = 1, \\
\frac{1}{\|\beta_{r,d}\|_2} \prod_{d=1}^{D} \|\beta_{r,d}\|_{2/D}^{1/D}, & \text{if } \lambda_2 = 0,
\end{cases}$$

10
where $\|\cdot\|_1$ is the $\ell_1$-norm of a vector. Further, with elastic-net penalty, the updating for $\tilde{B}_{D+1}$ in Algorithm 1 can be relaxed to a standard quadratically constrained quadratic program. Therefore, the dual ascent method and second-order cone programming can be used for this block-wise updating.

**Proposition 1.** Fix $\tilde{\nu}$ and $\tilde{B}_{D+1}$. Suppose $\Theta(\theta)$ is the scale class (modulo the sign) of

$$\theta = (B_1, \ldots, B_D)$$

up to scaling, i.e.,

$$\Theta(\theta) = \left\{ \theta^\rho : \theta^\rho = (B_1 \rho_1, \ldots, B_D \rho_D), \rho_d = \text{diag}(\rho_{1,d}, \ldots, \rho_{R,d}), \prod_{d=1}^D \rho_{r,d} = 1, \rho_{r,d} > 0, \text{ and } \rho_{r,d} = 1 \text{ if } \beta_{r,d} = 0 \right\},$$

where $\text{diag}(\rho_{1,d}, \ldots, \rho_{R,d}) \in \mathbb{R}^{R \times R}$ is a diagonal matrix with diagonal entries $\rho_{1,d}, \ldots, \rho_{R,d}$. Suppose (13) has a solution, denoted as $\hat{\rho}_{r,d}$, $r = 1, \ldots, R$, $d = 1, \ldots, D$. Let

$$\bar{\theta} = (\bar{B}_1, \ldots, \bar{B}_D),$$

where $\bar{B}_d = (\hat{\rho}_{1,d} \beta_{1,d}, \ldots, \hat{\rho}_{R,d} \beta_{R,d})$, $d = 1, \ldots, D$. Then

$$LG(\tilde{\nu}, \bar{\theta}, \tilde{B}_{D+1}) = \min_{\theta^\rho \in \Theta(\theta)} LG(\tilde{\nu}, \theta^\rho, \tilde{B}_{D+1}).$$

Furthermore, for the elastic-net penalty (14), if $\beta_{r,d} \neq 0$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, then

$$LG(\tilde{\nu}, \bar{\theta}, \tilde{B}_{D+1}) < LG(\tilde{\nu}, \theta^\rho, \tilde{B}_{D+1}), \quad \forall \theta^\rho \in \Theta(\theta), \quad \theta^\rho \neq \bar{\theta}.$$

Proposition 1 shows that $\bar{\theta}$ is the unique minimizer over $\Theta(\theta)$. This fixes the scaling indeterminacy, improves the practical convergence property and thus enhances the numerical performance. The numerical comparison between the algorithm with and without the rescaling strategy is presented in Section 5. Besides, the convergence of Algorithm 1 is shown in Proposition 2 and its proof is deferred to Section B.4 of the SM. Before presenting Proposition 2, we introduce an assumption on the penalty $G(\theta)$ as follows.

**Assumption 1.** $G(\theta)$ is a continuous function of $\theta$. There exists $\varpi(n) > 0$ (i.e., a sequence with respect to the sample size $n$) and a bounded constant $S_G > 0$ (independent of $\theta$ and $n$) such that

$$\varpi(n) \sum_{r=1}^R \sum_{d=1}^D \|\beta_{r,d}\|^2 \leq \{G(\theta)\}^{S_G}.$$
Further, denoting one accumulation point as an accumulation point of the sequence. Proposition 2.\(^{(\lambda)}\)

The knots are data-driven and chosen as equally spaced quantiles. As for \(K\), we follow Fan et al. (2014) to fix \(K = [2n^{4/5}]\), where \([\cdot]\) denotes rounding to the nearest integer. The knots are data-driven and chosen as equally spaced quantiles. As for \((R, \lambda_1, \lambda_2)\), cross-validation can be adopted. However, it is computationally expensive. For simplicity, we adopt the validation method in our numerical studies. The grids used in our experiments are given in Section 5.
4 Theoretical study

Throughout the theoretical analysis, we assume that the true regression function $m_0(X)$ has the following form of representation

$$m_0(X) = \nu_0 + \frac{1}{s} \sum_{r=1}^{R_0} \langle \beta_{0r,1} \circ \ldots \circ \beta_{0r,D}, F_0r(X) \rangle,$$

where $F_0r = B(f_0r, X)$ with $\int_0^{1} f_0r(x)dx = 0$, $f_0r \in \mathcal{H}$, $r = 1, \ldots, R_0$, and $\mathcal{H}$ is the function class specified in Assumption 4. In our analysis, we need the following regularity assumptions.

**Assumption 2.** The covariate tensor $X \in [0,1]^{p_1 \times \ldots \times p_D}$ has a continuous probability density function $g$, which is bounded away from zero and infinity on $[0,1]^{p_1 \times \ldots \times p_D}$, i.e., there exist constants $S_1, S_2 > 0$ such that $S_1 \leq g(x) \leq S_2$ for all $x \in [0,1]^{p_1 \times \ldots \times p_D}$.

Before presenting the assumption related to the random error, we first give the definition of sub-Gaussian random variable and its sub-Gaussian norm.

**Definition 1** (sub-Gaussian random variable). We say that a random variable $X$ is sub-Gaussian if there exists a positive constant $S$ such that

$$(\mathbb{E}|X|^p)^{1/p} \leq S \sqrt[p]{p}, \text{ for all } p \geq 1.$$

The minimum value of $S$ is the sub-Gaussian norm of $X$, denoted by $\|X\|_{\psi_2}$.

**Assumption 3.** The vector of random errors, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$, has independent and identically distributed entries. Each $\epsilon_i$ is sub-Gaussian with mean 0 and sub-Gaussian norm $\sigma < \infty$.

**Assumption 4.** The true broadcasted functions $f_0r \in \mathcal{H}$, $r = 1, \ldots, R_0$. Here $\mathcal{H}$ is the space of functions from $[0,1]$ to $\mathbb{R}$ satisfying the Hölder condition of order $\omega$, i.e.,

$$\mathcal{H} = \{g : |g^{(i)}(x_1) - g^{(i)}(x_2)| \leq \varsigma |x_1 - x_2|^\omega, \forall x_1, x_2 \in [0,1]\},$$

for some constant $\varsigma > 0$, where $i$ is a nonnegative integer and $g^{(i)}$ is the $i$-th derivative of $g$, such that $\omega \in (0, 1]$ and $\tau = i + \omega > 1/2$.

**Assumption 5.** The order of the B-spline used in (4) satisfies $\varsigma \geq \tau + \frac{1}{2}$. We let $0 = \xi_1 < \xi_2 < \cdots < \xi_{K-\varsigma+2} = 1$ denote the knots of B-spline basis and assume that

$$h_n = \max_{k=1,\ldots,K-\varsigma+1} |\xi_{k+1} - \xi_k| \asymp K^{-1} \quad \text{and} \quad h_n \left/ \min_{k=1,\ldots,K-\varsigma+1} |\xi_{k+1} - \xi_k| \right. \leq S_3,$$

for some constant $S_3 > 0$.

Assumptions 2, 4 and 5 are common in nonparametric regression models. In particular, Assumptions 4 and 5 regularize the space where the true broadcasted functions lie in and guarantee that they can be approximated well by B-spline functions. Indeed, it follows from these assumptions and Lemma 5 in Stone (1985) that there exist $\alpha_0, r = (\alpha_{0r,1}, \ldots, \alpha_{0r,K})^\top$, $r = 1, \ldots, R$, such that

$$\left\| f_{0r} - \sum_{k=1}^{K} \alpha_{0r,k} b_k \right\|_\infty = \mathcal{O}(K^{-\tau}), \quad (15)$$
where the $L_\infty$-norm of a uni-dimensional function $f$ is defined as $\|f\|_\infty = \sup_x |f(x)|$. Although we assume $\int_0^1 f_{or}(x) dx = 0$, Lemma A.6 (in the SM) still implies that there are $\alpha_{0,r}$, $r = 1, \ldots, R$, satisfying (15) with

$$\sum_{k=1}^K \int_0^1 \alpha_{0r,k} b_k(u) du = 0. \quad (16)$$

Despite this mild difference in parameter identification, similar assumptions can be found in Zhou et al. (1998) and Huang et al. (2010). Besides, the sub-Gaussianity condition in Assumption 3 is now a standard tail condition of the error.

We present the convergence rates of $\hat{m}_{\text{PLS}}(X)$. The function norm for a tensor function $m$ is defined as $\|m\| = \{\mathbb{E}_X m^2(X)\}^{1/2}$, which is equivalent to the $L_2$-norm, $\|m\|_{L_2} = \{\int m^2(X)dX\}^{1/2}$, due to Assumption 2. To simplify the notations, we write $B_{0r} = \beta_{0r,1} \cdots \beta_{0r,D}$. Theorem 1 is the major theorem that provides a non-asymptotic error upper bound for the general form of the proposed estimator, where $p_i$, $K$, $R$ and $R_0$ are allowed to go to infinity with the sample size $n$. In Corollaries 1 and 2, we will degenerate Theorem 1 to two special cases which are easier to comprehend.

**Theorem 1.** Let $\hat{m}_{\text{PLS}}(X)$ be the estimated regression function reconstructed from (10). If Assumptions 1–5 hold, $R \geq R_0$ and $n \geq C_1 h_n^{-2 - 2/\log h_n}(\log^{-2} h_n)(\Delta + \sum_{i=1}^D R P_i + RK)$, for some large enough constant $C_1 > 0$, then

$$\|\hat{m}_{\text{PLS}} - m_0\|^2 \leq C_2 \left\{ \frac{\Delta + \sum_{i=1}^D R P_i + RK}{n} \right\} + C_3 \left\{ \sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1 \right\}^2 \frac{1}{K^{2r}} + C_4 \frac{G_0}{n}, \quad (17)$$

with probability at least

$$1 - C_5 \exp \left\{ - C_6 \left( \Delta + R \sum_{i=1}^D P_i + RK \right) \right\},$$

where $C_2, \ldots, C_6 > 0$ are some constants, $G_0$ is defined in (A.9) of the SM, and

$$\Delta = \min\{ R^{D+1}, R^2 \log \delta_{\text{pen}} \}. \quad (18)$$

Here, if one adopts a penalty (see Corollary 2 for an example) such that the bias term $G_0/n$ is dominated by the previous two terms in the right hand side of (17), then

$$\log \delta_{\text{pen}} = O \left( \log \left( \max \left\{ n, \frac{1}{\omega(n)}, \max \{ \beta_{0r,d,l}, \max \{ \alpha_{0r,k}, \nu_0 \} \} \right\} \right) \right).$$

The explicit definition of $\delta_{\text{pen}}$ is shown in (A.26) of the SM.

Theorem 1 involves the complicated quantities $G_0$ and $\delta_{\text{pen}}$ that are specified in the SM. Note that, from (18), $\Delta$ is upper bounded by both $R^{D+1}$ and $R^2 \log \delta_{\text{pen}}$. Ignoring $R^2 \log \delta_{\text{pen}}$, our theorem still guarantees a bound for $\Delta$ and so the error of the proposed estimator. We will demonstrate that $R^2 \log \delta_{\text{pen}}$ leads to a better bound in a wide range of settings. For the unpenalized estimator, one can show that $G_0 = 0$ and $\Delta = R^{D+1}$. As a solid special case of Theorem 1, the following Corollary 1 explicitly states the corresponding results for the unpenalized estimator.
Corollary 1. Suppose \( \hat{m}_{\text{LS}}(X) \) is the estimated regression function reconstructed from (7). If Assumptions 2–5 hold, \( R \geq R_0 \), and \( n > C_7 h_n^{-2-2/(\log h_n)}(\log^{-2} h_n)\left(R^{D+1} + \sum_{i=1}^{D} R p_i + R K\right) \) for some large enough constant \( C_7 > 0 \), then we have the following result:

\[
\| \hat{m}_{\text{LS}} - m_0 \|^2 = O_p \left( \frac{R^{D+1} + \sum_{i=1}^{D} R p_i + R K}{n} \right) \\
+ O_p \left( \left\{ \frac{\sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1}{s} \right\}^2 \frac{1}{K^{2\tau}} \right). \tag{19}
\]

The proof of Theorem 1 is not straightforward. To see this, if we discard the low-rank and broadcasting structure of the proposed model (3), we can rewrite the regression function as a nonparametric additive model, and vectorize the basis tensor and its coefficients in (5), i.e.,

\[
m(X_i) = \nu + z_i^T a,
\]

where \( z_i = \text{vec}\{\Phi(X_i)\} \) and \( a = \text{vec}(A) \). The main challenge of studying the convergence rates is to determine the upper and lower bounds for the eigenvalues of the Gram matrix of “design”, i.e., \( Z^T Z/n \), where \( Z = (z_1, \ldots, z_n)^T \). For a fixed number of predictors, Huang et al. (2010) shows the bounds of the eigenvalues using Lemma 3 of Stone (1985) and Lemma 6.2 of Zhou et al. (1998). It is worth mentioning that directly using the results of Stone (1985) will result in a diminishing lower bound at an exponential rate of \( s \), when the number of predictors \( s \) goes to infinity with the sample size \( n \) (see, e.g., Chen et al., 2018). Therefore, a new study of eigenvalue bounds is needed to carefully harness the model structure, in particular the low-CP-rank structure of \( A \) (see (7)). In our proof, we obtain well-controlled bounds of the restricted eigenvalues over a set of low-CP-rank tensors, that holds with high probability when the sample size is of order \( K^2 \{ \min(R^{D+1}, R^2 \log \delta_{\text{pen}}) + \sum_{i=1}^{D} R p_i + R K \} \); see the more precise version (A.16), (A.25), and (A.81) in the SM. The resulting eigenvalue bounds fill in the gap, with a reasonable sample size dependence. Besides, in order to derive a potentially tighter bound \( (R^2 \log \delta_{\text{pen}}) \) in \( \Delta \), we have also developed a novel entropy result (Lemma A.8 in the SM).

Suppose we choose a penalty to make the bias term to be dominated by other terms in the right hand side of (17). Roughly speaking, the first and the second terms in (19) correspond to the estimation error and the approximation error, respectively. We can see that the estimation error roughly scales with the number of effective parameters in the model. For different combinations of orders between the parameters \( (R, R_0, p_i) \) and the sample size \( n \), we can tune the number of basis functions \( K \) to achieve the best rate of convergence. Let

\[
\delta_1 = \Delta + \sum_{i=1}^{D} R p_i \quad \text{and} \quad \delta_2 = \left\{ \frac{\sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1}{s} \right\}^2.
\]

If \( n^{1/(2\tau+1)} = O(\delta_1 \delta_2^{-1/(2\tau+1)} R^{-2\tau/(2\tau+1)}) \), the best rate is \( \delta_1/n \) when \( K \) satisfies

\[
(n\delta_2/\delta_1)^{1/2\tau} \lesssim K \lesssim \delta_1/R,
\]

where \( a \lesssim b \) also means \( a = O(b) \). On the other hand, if \( \delta_1 \delta_2^{-1/(2\tau+1)} R^{-2\tau/(2\tau+1)} = o(n^{1/(2\tau+1)}) \), letting \( K \propto (n\delta_2/R)^{1/(2\tau+1)} \) results in the best rate \( (R/n)^{2\tau/(2\tau+1)} \delta_2^{1/(2\tau+1)} \).
One special case is that when $p_i, R$ and $R_0$ do not grow with $n$, choosing $K \propto n^{1/(2\tau+1)}$ leads to the optimal rate of convergence $n^{-2\tau/(2\tau+1)}$ as in Stone (1982). Theorem 1 indeed generalizes the canonical results to tensor low-rank modeling with broadcasting.

Note that Theorem 1 involves the term $\delta_{\text{pen}}$ when penalization is effective. As explained before, this could lead to a tighter bound for $\Delta$ and so the convergence rate of the proposed estimator. While the explicit interplay between different terms are detailed in Theorem 1, we provide an example in Corollary 2 to demonstrate this. Denote

$$
\mathcal{M}_{00} = \left\{ m(X) : m(X) = \nu + \sum_{r=1}^R \left< \beta_{r,1}, \ldots, \beta_{r,D} \right>_p F_r(X) \right\},
$$

where

$$
\nu \leq V_1, \left\| \beta_{r,d} \right\|_1 \leq V_2, \beta_{r,d} \in \mathbb{R}^p, \int_0^1 f_r^2(t) \, dt \leq V_3,
$$

(20)

where $V_1, V_2, V_3 > 0$ are some constants.

**Corollary 2.** Suppose the same conditions of Theorem 1 hold. Assume $m_0 \in \mathcal{M}_{00}$. If we adopt the ridge penalty (i.e., (12) with (14) and $\lambda_2 = 0$) and choose a $\lambda_1$ that satisfies (A.28) in the SM, we then have $\log \delta_{\text{pen}} \leq V_4 K \log n$, where $V_4 > 0$ is a constant, and the bias term is dominated by its preceding two terms. Thus, (17) can be written as

$$
\| \hat{m} - m_0 \|^2 \leq V_5 \left\{ \min(R^{D+1}, R^2 K \log n) + \sum_{i=1}^D R p_i + RK \right\} + V_6 \frac{R^2}{K^{2\tau}},
$$

where $V_5, V_6 > 0$ are some constants.

The proof of Corollary 2 is deferred to Section A.4 of the SM. One can obtain similar result for our estimator combined with elastic-net penalty for any fixed $\lambda_2 \in [0, 1]$, and the corresponding error upper bounds have the same order as the one shown in Corollary 2. Further improvement and discussions are deferred to Section A.7 of the SM.

To evaluate the optimality of our upper bounds, we provide a minimax lower bound as follow.

**Theorem 2.** Suppose the random error $\epsilon$ is a Gaussian random variable with variance $\sigma^2 < \infty$. Under the Assumptions 2 and 4, if $\max_d p_d \geq V_7 R$, where $V_7 > 0$ is a constant, we have

$$
\inf_{\hat{m}} \sup_{m \in \mathcal{M}_{00}} \mathbb{E}(\| \hat{m} - m \|^2) \geq S_4 \max \left\{ \min \left( \sum_d p_d R, S_5 \right), \left( \frac{R}{n} \right)^{2\tau/(2\tau+1)} \right\},
$$

(21)

where $S_4, S_5 > 0$ are some constants and $\mathcal{M}_{00}$ is defined in (20).

The proof of Theorem 2 is deferred to Section A.5 in the SM. First, we discuss the general case where $R_0$ is allowed to grow with the sample size $n$. Suppose $R \geq R_0$ and $R \simeq R_0$. When $\sum_d p_d R/n$ dominate $(R/n)^{2\tau/(2\tau+1)}$ in (21), the upper bound in Theorem 1 can match the minimax lower bound in terms of order, based on mild conditions. To be more specific, assume the size of the tensor satisfies

$$
\max \left[ (R_0 n)^{1/(2\tau+1)}, \min \{ R_0^D, R_0 \log \delta_{\text{pen}} \} \right] \lesssim \sum_d p_d \lesssim \frac{n^{(2\tau-1)/(2\tau+1)} \log^3 n}{R_0 \log n - R_0(2\tau + 1)},
$$

(22)
Then, with proper choice of $K$ and penalty such that $G_0/n$ is dominated by the first two terms in the right hand side of (17), our estimator achieves the minimax optimal rate. For the elastic-net penalty with a fixed $\lambda_2 \in [0,1]$ and an appropriate choice of $\lambda_1$, the lower bound in the Condition (22) can be relaxed to

$$\min \left[ \max \{ (R_0 n)^{-\tau+1}, R_0^D \}, R_0 n^{-\tau+1} (\log n)^{2\tau+1} \right].$$

(23)

On the other hand, if $R_0$ is bounded, then the Condition (22) can be replaced by

$$\sum_d P_d = O\left( n^{(2\tau-1)/(2\tau+1)} \log^3 n \right).$$

Further discussions on the lower bound are deferred to Section A.7 of the SM.

Finally, we remark that there exists a gap between the estimators studied in Proposition 2 and Theorem 1 respectively. Theorem 1 gives the property of the global optimum of (10), whereas Proposition 2 showed that any accumulation point of the sequence generated by Algorithm 1 is a stationary point, which is not necessarily the global optimum. However, our empirical experiments confirm that the output of the proposed algorithm has good performance, and indicates that an essential theoretical property that bridges the gap tends to hold with high probability in practice. Further theoretical insights and empirical exploration about this gap is presented in Section F of the SM.

5 Experiments

To evaluate the empirical performance of the proposed broadcasted nonparametric tensor regression (BroadcastTR) with the elastic-net penalty, we compared BroadcastTR with two alternatives upon both synthetic (Section 5.1) and real data (Section 5.2) sets. These alternatives are (i) elastic-net regression on the vectorized tensor predictor (ENetR) (Zou and Hastie, 2005) and (ii) tensor linear regression (TLR) (Zhou et al., 2013). Throughout the numerical experiments, ENetR and TLR were implemented by the R package “glmnet” (Friedman et al., 2010) and the MATLAB toolbox “TensorReg” (Zhou et al., 2013) respectively. Since the proposed rescaling strategy (13) can be applied to the computation of tensor linear regression, we also considered this algorithmic modification in our study. To distinguish this modification, we use TLR and TLR-rescaled to represent the algorithm of Zhou et al. (2013) and our algorithm with scaling strategy respectively. We also provided more numerical experiments in the supplementary material. Specifically, an additional real data analysis, based on a neuroimaging dataset from Alzheimer’s Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005) is presented in Section G of the SM. Further numerical comparisons with existing nonlinear tensor regression models, including tensor-variate Gaussian process regression (TVGP, Zhao et al., 2014) and Gaussian process nonparametric tensor estimator (GPNTE, Kanagawa et al., 2016) are provided in Section H.

We aim to evaluate estimation and prediction performance as well as region selection of these methods. To identify entry-wise contribution of the covariate tensor, we note that the estimated regression function $\hat{m}$ of the above methods can be expressed as an additive form $\hat{\nu} + \sum_{i_1,...,i_D} \hat{m}_{i_1,...,i_D}(X_{i_1,...,i_D})$, where $\int_0^1 \hat{m}_{i_1,...,i_D}(x)dx = 0$, and so the entry-wise effect can be summarized by the $L_2$-norm $\| \hat{m}_{i_1,...,i_D} \|_{L_2} = \left\{ \int_0^1 \hat{m}_{i_1,...,i_D}^2(x)dx \right\}^{1/2}$. More specifically,
Putting together these entrywise nonlinear effects, we obtain a tensor of dimension $p_1 \times \cdots \times p_D$ with the $(i_1, \ldots, i_D)$-th element being $||\hat{m}_{i_1, \ldots, i_D}||_{L_2}$. In below, it is called the norm tensor of the corresponding tensor regression method. We used this norm tensor to indicate important subregions identified by all methods. The tuning parameters were selected as in Section 3. For the grids of $R$ and $(\lambda_1, \lambda_2)$, we followed the suggestions of He et al. (2018); Engebretsen and Bohlin (2019); Teipel et al. (2015). In particular, the following grids were considered in the synthetic experiments: $R \in \{1, 2, 3, 4, 5\}$, $\lambda_1 \in \{10^{-2}, 5 \times 10^{-1}, 10^{-1}, \ldots, 10^{2}, 5 \times 10^{2}, 10^{3}\}$ and $\lambda_2 \in \{0, 0.5, 1\}$, for BroadcasTR, TLR, and TLR-rescaled. As for ENetR, we used the same grids of $\lambda_1$ and $\lambda_2$. In the real data analysis, due to its sample size is larger, we considered the following grids for the rank $R \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, and penalized parameters $\lambda_1 \in \{10^{-2}, 5 \times 10^{-2}, 5 \times 10^{-2}, 7.5 \times 10^{-2}, 10^{-1}, \ldots, 10^{2}, 2.5 \times 10^{2}, 5 \times 10^{2}, 7.5 \times 10^{2}, 10^{3}\}$, $\lambda_2 \in \{0, 0.5, 1\}$ for BroadcasTR, TLR, TLR-rescaled, and the same grids of $\lambda_1$ and $\lambda_2$ for ENetR. We find that these grids are wide enough in our experiments in the sense that the boundary grid points are seldomly selected.

5.1 Synthetic data

Similar to Zhou et al. (2013), we fix the dimension of $X$ to be $64 \times 64$. In our simulation study, we consider four different regression functions:

Case 1: $y = m_1(X) + \epsilon_1 = 1 + \langle B_1, X \rangle + \epsilon_1,$

Case 2: $y = m_2(X) + \epsilon_2 = 1 + \langle B_2, F_1(X) \rangle + \epsilon_2,$

Case 3: $y = m_3(X) + \epsilon_3 = 1 + \langle B_3, F_1(X) \rangle + \epsilon_3,$

Case 4: $y = m_4(X) + \epsilon_4 = 1 + \langle B_{41}, F_1(X) \rangle + \langle B_{42}, F_2(X) \rangle + \epsilon_4,$

where $F_1, F_2 : [0, 1]^{64 \times 64} \rightarrow \mathbb{R}^{64 \times 64}$ satisfy

$$(F_1(X))_{i_1, i_2} = f_1(X_{i_1, i_2}) = X_{i_1, i_2} + 0.6 \sin\{2\pi(X_{i_1, i_2} - 0.5)^2\},$$

$$(F_2(X))_{i_1, i_2} = f_2(X_{i_1, i_2}) = X_{i_1, i_2} + 0.3 \cos(2\pi X_{i_1, i_2}),$$

for $i_1 = 1, \ldots, 64$ and $i_2 = 1, \ldots, 64$. The scaling matrices $B_1, B_2, B_3, B_{41}$ and $B_{42}$ are binary and depicted in the first column of Figure 4. These regression functions are used to illustrate four different situations: (1) a linear model with one important rank-two subregion; (2) a nonlinear model with one important rank-four subregion; (3) a nonlinear model with two separated important rank-two subregions that share the same nonlinearity; and (4) a low rank nonlinear model with two separated important rank-two subregions that show different nonlinearities.

For each Case $j$, the covariate tensor $X$ and the error $\epsilon_j$ were generated such that $X_{i_1, i_2} \sim \text{Uniform}[0, 1]$, $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ independently across all $i_1, i_2$. The parameter $\sigma_j$ was set to be 10% of the standard deviation of entries of $m_j(X)$. We generated 50 simulated data sets independently for each setting of sample size $n = 500, 750, 1000$. Each simulated data set was then split into two separate subsets: a training set with 80% data and a validation set with
The tuning parameters of the underlying methods were selected by minimizing the validation error:

$$\frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} (y_{\text{valid},i} - \hat{y}_{\text{valid},i})^2,$$

over grids of corresponding tuning parameters, where $n_{\text{valid}}$ is the size of the validation set, $\hat{y}_{\text{valid},i}$ is the prediction value of the $i$-th observation $y_{\text{valid},i}$ in the validation set.

To evaluate the estimation performance, we define the integrated squared error (ISE)

$$\text{ISE} = \|\hat{m} - m_0\|_{L_2}^2,$$

where $m_0$ and $\hat{m}$ are the true function and a generic estimated function respectively. We note that $\|\hat{m} - m_0\|_{L_2} = \|\hat{m} - m_0\|$ (which we adopt in Section 4), due to the uniform distribution of $X$. The average ISEs of the proposed and comparative methods are summarized in Table 1. For the nonlinear situations, i.e., Cases 2–4, it is shown that the proposed BroadcasTR outperforms the other methods significantly. In particular, BroadcasTR reduces the average ISEs by 86%–99% in Case 2, 65%–99% in Case 3, and 74%–98% in Case 4. As for Case 1, which is the linear setting and in favor of the alternative methods, BroadcasTR remains competitive. It performs better than both TLR and ENetR by showing 54%–99% reduction in the average ISEs, and is slightly inferior to TLR-rescaled. Besides, TLR-rescaled performs much better than TLR although they originate from the same penalized regression. This indicates that the proposed rescaling strategy leads to significant improvements. Furthermore, the accuracy of estimation increases with the sample size for the proposed BroadcasTR, which is consistent with our asymptotic analysis.

The important subregions for BroadcasTR, TLR, TLR-rescaled, and ENetR are identified by their norm tensors (matrices) as defined above. For each method, the norm tensor with the median ISE among 50 simulated datasets of $n = 1000$ was depicted in Figure 4. It shows that BroadcasTR, TLR and TLR-rescaled have similar region selection result for Case 1 (a low-rank model with linear effects), whereas BroadcasTR is much better than TLR and TLR-rescaled for Cases 2–4 (low-rank models with nonlinearity). In all cases, ENetR is unable to identify the import regions, which is an empirical evidence that incorporating the tensor low-rank structure can improve region selection, hence enhancing interpretability. We also present the region identification performance of BroadcasTR for smaller sample sizes ($n = 500$ or 750) in Figure 5. It is not surprising to see that when the sample size increases, the accuracy of identified regions of our proposed method improves.

5.2 Monkey’s electrocorticography data

We also evaluated various methods on a publicly available monkey’s electrocorticography (ECoG) data set (Shimoda et al., 2012). The corresponding tensor covariate is a preprocessed ECoG signal (Shimoda et al., 2012), organized as a third order tensor of dimensions $64 \times 10 \times 10$ (channel × frequency × time), and the response variable is the movement distance of a monkey’s left shoulder marker along a particular direction. The data preprocessing procedure is similar to Shimoda et al. (2012) which tracks 15-minute experiments. Corresponding details are given in Section E of the SM. Following Hou et al. (2015), we chose 10000 observations of the whole data set starting from the second minute of the experiments. The data set was
Table 1: Estimation performance in synthetic data. Reported are the averages of ISEs and its standard deviation (in parenthesis) based on 50 data replications. In the first column, \( n \) is the total sample size, of which 20% were kept for validation.

| \( n \) | Case | TLR     | TLR-rescaled | ENetR     | BroadcasTR |
|-------|------|---------|-------------|----------|------------|
| 500   | 1    | 0.251 (0.054) | **0.066** (0.012) | 16.559 (0.722) | 0.092 (0.018) |
|       | 2    | 24.422 (2.531) | 22.282 (1.683) | 31.255 (1.261) | **3.185** (1.843) |
|       | 3    | 76.331 (5.371) | 71.915 (6.543) | 75.425 (1.770) | **25.228** (5.412) |
|       | 4    | 90.285 (7.081) | 89.560 (6.742) | 89.253 (3.060) | **23.647** (6.330) |
| 750   | 1    | 0.115 (0.024) | **0.038** (0.007) | 14.747 (0.606) | 0.053 (0.009) |
|       | 2    | 20.899 (1.444) | 17.039 (1.665) | 30.599 (0.897) | **0.640** (0.186) |
|       | 3    | 71.609 (9.236) | 53.298 (4.244) | 73.936 (2.681) | **3.373** (2.358) |
|       | 4    | 80.444 (12.828) | 57.946 (4.459) | 86.782 (3.576) | **4.166** (2.807) |
| 1000  | 1    | 0.077 (0.013) | **0.028** (0.007) | 10.122 (1.021) | 0.038 (0.006) |
|       | 2    | 18.815 (2.133) | 15.212 (0.969) | 29.718 (0.637) | **0.258** (0.056) |
|       | 3    | 65.387 (10.013) | 45.331 (2.868) | 71.729 (2.130) | **0.705** (0.124) |
|       | 4    | 63.848 (10.967) | 51.381 (2.583) | 84.549 (2.355) | **1.046** (0.177) |

Table 2: Prediction performance on the monkey’s electrocorticography data. Reported are averages of MSPE and its standard deviation (in parenthesis) based on 10 random splittings.

| Data   | TLR     | TLR-rescaled | ENetR     | BroadcasTR |
|--------|---------|-------------|----------|------------|
| Monkey | 3.1703 (0.0418) | 3.0923 (0.0699) | 3.1256 (0.0431) | **2.5468** (0.0961) |

then randomly split into three different subsets, i.e., a training set, a validation set, and a test set, of size 4000, 1000, and 5000 respectively.

To measure the performance, we use mean squared prediction error (MSPE)

\[
\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_{\text{test},i} - \hat{y}_{\text{test},i})^2, \quad (24)
\]

where \( n_{\text{test}} \) is the size of the test set, \( \hat{y}_{\text{test},i} \) is the prediction value of the \( i \)-th observed value \( y_{\text{test},i} \) in the test set. We repeated the fittings for 10 random splittings. The average MSPE over these 10 fittings are reported in Table 2, from which we can see BroadcasTR performs significantly better than the others in prediction.

6 Conclusion

In this paper, we have proposed a broadcasted model to study the problem of nonlinear regressions with tensor covariates. The curse of dimensionality is tamed by simultaneously utilizing the low-rank tensor structure and broadcasting a uni-dimensional function within each component. With a regularized estimation, the proposed model shows the advantages of improved prediction performance and identifying the important regions on the tensor covariates. Moreover, the convergence rates of the estimator are derived, based on a novel restricted
Figure 4: Region selection of TLR, TLR-rescaled, ENetR, and BroadcasTR for $n = 1000$, of which 20% were for validation. The first column presents the true scaling tensors, which are $B_1$, $B_2$, $B_3$ and $B_{41} + B_{42}$ for Cases 1, 2, 3 and 4, respectively. The rest four columns depict the estimated norm tensor with median ISEs of the comparative and proposed methods. Columns from left to right respectively correspond to TLR, TLR-rescaled, ENetR, and BroadcasTR. The plots in all columns share the same color scheme as shown in the color bar at the bottom.

eigenvalue result. We use both synthetic and real data sets to evaluate the empirical performance of the proposed broadcasted nonparametric regression model with some comparison methods, and the results confirm our theoretical findings. The convergence of our proposed estimation algorithm can be guaranteed, with the proposed generalized rescaling strategy. Although our method has concentrated on the problem of continuous and univariate response throughout the paper, it is not difficult to generalize it to a classification paradigm or models with multivariate responses. When one dimension of the tensor covariates is ultra-high, using the low-rank structure alone may not be sufficient to obtain a consistent estimation with promising prediction performance. Thus one interesting future research topic is to develop an alternative method incorporating the low-rank tensor with the entry-wise or slice-wise sparsity. Moreover, modeling the regression function by index models with dimension reduction
Figure 5: Region selection of BroadcasTR for Cases 1–4, with various sample size $n = 500, 750, 1000$ (where 20% data are used for validation). All plots share the same color scheme as shown in the color bar at the bottom.

techniques of tensor covariates is also of interest, and needs further investigation.

References

Absil, P. A., Mahony, R. and Sepulchre, R. (2008) Optimization Algorithms on Matrix Manifolds. Princeton University Press.

Alquier, P. and Biau, G. (2013) Sparse single-index model. Journal of Machine Learning Research, 14, 243–280.

Banerjee, A., Chen, S., Fazayeli, F. and Sivakumar, V. (2015) Estimation with norm regularization. arXiv preprint arXiv:1505.02294.
de Boor, C. (1973) The quasi-interpolant as a tool in elementary polynomial spline theory. In *Approximation Theory* (ed. G. G. Lorentz), 269–276. Academic Press.

— (1976) Splines as linear combinations of b-splines: A survey. *Tech. rep.*, Wisconsin University Madison Mathematics Research Center.

de Boor, C. and Fix, G. (1973) Spline approximation by quasiinterplonts. *Journal of Approximation Theory*, 8, 19–45.

Chandrasekaran, V., Recht, B., Parrilo, P. A. and Willsky, A. S. (2012) The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12, 805–849.

Chao, Z. C., Nagasaka, Y. and Fujii, N. (2010) Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey. *Frontiers in Neuroengineering*, 3, 3.

Chen, H., Raskutti, G. and Yuan, M. (2019) Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20, 172–208.

Chen, Z., Fan, J. and Li, R. (2018) Error variance estimation in ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 113, 315–327.

Davatzikos, C., Vaillant, M., Resnick, S. M., Prince, J. L., Letovsky, S. and Bryan, R. N. (1996) A computerized approach for morphological analysis of the corpus callosum. *Journal of Computer Assisted Tomography*, 20, 88–97.

De Lathauwer, L. (2006) A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 28, 642–666.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000) A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21, 1253–1278.

Durham, T. J., Libbrecht, M. W., Howbert, J. J., Bilmes, J. and Noble, W. S. (2018) Predicted parallel epigenomics data imputation with cloud-based tensor decomposition. *Nature Communications*, 9, 1402.

Engebretsen, S. and Bohlin, J. (2019) Statistical predictions with glmnet. *Clinical Epigenetics*, 11, 1–3.

Fan, J., Feng, Y. and Song, R. (2011) Nonparametric independence screening in sparse ultrahigh-dimensional additive models. *Journal of the American Statistical Association*, 106, 544–557.

Fan, J., Ma, Y. and Dai, W. (2014) Nonparametric independence screening in sparse ultrahigh-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109, 1270–1284.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.
Hao, B., Wang, B., Wang, P., Zhang, J., Yang, J. and Sun, W. W. (2019) Sparse tensor additive regression. arXiv preprint arXiv:1904.00479.

Harshman, R. A. (1984) Data preprocessing and the extended parafac model. Research Methods for Multi-mode Data Analysis, 216–284.

Hastie, T. J. and Tibshirani, R. J. (1990) Generalized additive models. London: CRC Press.

He, L., Wang, F., Chen, K., Xu, W. and Zhou, J. (2018) Boosted sparse and low-rank tensor regression. In Advances in Neural Information Processing Systems, 1009–1018.

Hlavácková-Schindler, K. (2010) A new lower bound for the minimal singular value for real non-singular matrices by a matrix norm and determinant. Applied Mathematical Sciences, 4, 3189–3193.

Hoff, P. D. (2015) Multilinear tensor regression for longitudinal relational data. The Annals of Applied Statistics, 9, 1169–1193.

Horowitz, J. L. and Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. Journal of the American Statistical Association, 91, 1632–1640.

Hou, M., Wang, Y. and Chaib-draa, B. (2015) Online local gaussian process for tensor-variate regression: Application to fast reconstruction of limb movements from brain signal. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, 5490–5494. IEEE.

Hu, W., Kong, D. and Shen, W. (2019) Nonparametric matrix response regression with application to brain imaging data analysis. arXiv preprint arXiv:1904.00495.

Huang, J., Horowitz, J. L. and Wei, F. (2010) Variable selection in nonparametric additive models. The Annals of Statistics, 38, 2282–2313.

Huang, X. and Pan, W. (2003) Linear regression and two-class classification with gene expression data. Bioinformatics, 19, 2072–2078.

Ichimura, H. (1993) Semiparametric least squares (sls) and weighted sls estimation of single-index models. Journal of Econometrics, 58, 71–120.

Imaizumi, M. and Hayashi, K. (2016) Doubly decomposing nonparametric tensor regression. In International Conference on Machine Learning, 727–736.

Kanagawa, H., Suzuki, T., Kobayashi, H., Shimizu, N. and Tagami, Y. (2016) Gaussian process nonparametric tensor estimator and its minimax optimality. In International Conference on Machine Learning, 1632–1641.

Kang, J., Reich, B. J. and Staicu, A.-M. (2018) Scalar-on-image regression via the soft-thresholded gaussian process. Biometrika, 105, 165–184.

Khosla, M., Jamison, K., Kuceyeski, A. and Sabuncu, M. R. (2018) 3d convolutional neural networks for classification of functional connectomes. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 137–145. Springer.
Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *SIAM Review*, 51, 455–500.

Koltchinskii, V. (2011) *Oracle inequalities in empirical risk minimization and sparse recovery problems*. New York: Springer Science & Business Media.

Kruskal, J. B. (1977) Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18, 95–138.

— (1989) Rank, decomposition, and uniqueness for 3-way and n-way arrays. *Multiway Data Analysis*, 7–18.

Li, L. and Zhang, X. (2017) Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112, 1131–1146.

Li, X., Morgan, P. S., Ashburner, J., Smith, J. and Rorden, C. (2016) The first step for neuroimaging data analysis: Dicom to nifti conversion. *Journal of Neuroscience Methods*, 264, 47–56.

Li, X., Xu, D., Zhou, H. and Li, L. (2018) Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10, 520–545.

Lin, Y. and Zhang, H. H. (2006) Component selection and smoothing in multivariate non-parametric regression. *The Annals of Statistics*, 34, 2272–2297.

Lock, E. F. (2018) Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics*, 27, 638–647.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R. and Beg, M. F. (2018) Multiscale deep neural networks based analysis of fdg-pet images for the early diagnosis of alzheimer’s disease. *Medical Image Analysis*, 26–34.

Lu, H., Plataniotis, K. N. and Venetsanopoulos, A. (2013) *Multilinear subspace learning: Dimensionality reduction of multidimensional data*. Boca Raton, Florida: Chapman and Hall/CRC.

McFarland, D. J., McCane, L. M., David, S. V. and Wolpaw, J. R. (1997) Spatial filter selection for eeg-based communication. *Electroencephalography and Clinical Neurophysiology*, 103, 386–394.

Meier, L., Van De Geer, S. and Bühlmann, P. (2009) High-dimensional additive modeling. *The Annals of Statistics*, 37, 3779–3821.

Miranda, M. F., Zhu, H. and Ibrahim, J. G. (2018) Tprm: Tensor partition regression models with applications in imaging biomarker detection. *The Annals of Applied Statistics*, 12, 1422–1450.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W. and Beckett, L. (2005) The alzheimer’s disease neuroimaging initiative. *Neuroimaging Clinics*, 15, 869–877.
Radchenko, P. (2015) High dimensional single index models. *Journal of Multivariate Analysis, 139*, 266–282.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. and Shulman, G. L. (2001) A default mode of brain function. *Proceedings of the National Academy of Sciences, 98*, 676–682.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research, 13*, 389–427.

Raskutti, G., Yuan, M. and Chen, H. (2019) Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics, 47*, 1554–1584.

Rauhut, H., Schneider, R. and Stojanac, Ž. (2017) Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications, 523*, 220–262.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*, 1009–1030.

Reiss, P. T. and Ogden, R. T. (2010) Functional generalized linear models with images as predictors. *Biometrics, 66*, 61–69.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric regression*. Cambridge: Cambridge University Press.

Salat, D. H., Kaye, J. A. and Janowsky, J. S. (2001) Selective preservation and degeneration within the prefrontal cortex in aging and alzheimer disease. *Archives of Neurology, 58*, 1403–1408.

Selvan, S. E., Amato, U., Gallivan, K. A., Qi, C., Carfora, M. F., Larobina, M. and Alfano, B. (2012) Descent algorithms on oblique manifold for source-adaptive ica contrast. *IEEE Transactions on Neural Networks and Learning Systems, 23*, 1930–1947.

Shimoda, K., Nagasaka, Y., Chao, Z. C. and Fujii, N. (2012) Decoding continuous threedimensional hand trajectories from epidural electrocorticographic signals in japanese macaques. *Journal of Neural Engineering, 9*, 036015.

Sidiropoulos, N. D. and Bro, R. (2000) On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics: A Journal of the Chemometrics Society, 14*, 229–239.

Smilde, A., Bro, R. and Geladi, P. (2005) *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W. and Schacter, D. L. (2010) Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *Neuroimage, 53*, 303–317.

Stegeman, A. and Sidiropoulos, N. D. (2007) On kruskal’s uniqueness condition for the candecomp/parafac decomposition. *Linear Algebra and its Applications, 420*, 540–552.
Stone, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053.

— (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, 3, 689–705.

Sun, W. W. and Li, L. (2017) Store: Sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18, 4908–4944.

Suzuki, T. (2015) Convergence rate of bayesian tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, 1273–1282.

Talagrand, M. (2005) *The generic chaining*. Berlin: Springer.

Teipel, S. J., Bayer, W., Alexander, G. E., Zebuhr, Y., Teichberg, D., Kulic, L., Schapiro, M. B., Möller, H.-J., Rapoport, S. I. and Hampel, H. (2002) Progression of corpus callosum atrophy in alzheimer disease. *Archives of Neurology*, 59, 243–248.

Teipel, S. J., Kurth, J., Krause, B., Grothe, M. J., Initiative, A. D. N. et al. (2015) The relative importance of imaging markers for the prediction of alzheimer’s disease dementia in mild cognitive impairment—beyond classical regression. *NeuroImage: Clinical*, 8, 583–593.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

de la Torre, J. C. (2010) Alzheimer’s disease is incurable but preventable. *Journal of Alzheimer’s Disease*, 20, 861–870.

Vershynin, R. (2018) *High-dimensional probability: An introduction with applications in data science*. New York: Cambridge University Press.

Wang, F., Zhang, P., Qian, B., Wang, X. and Davidson, I. (2014) Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 145–154. ACM.

Wang, X., Zhu, H. and Initiative, A. D. N. (2017) Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112, 1156–1168.

Wood, S. N. (2017) *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.

Yang, Y. and Barron, A. (1999) Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 1564–1599.

Yang, Y. and Tokdar, S. T. (2015) Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43, 652–674.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.
Zhang, A. R., Luo, Y., Raskutti, G. and Yuan, M. (2020) Islet: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science, 2*, 444–479.

Zhang, B., Zhou, H., Wang, L. and Sung, C. (2017) Classification based on neuroimaging data by tensor boosting. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 1174–1179. IEEE.

Zhang, X., Li, L., Zhou, H., Zhou, Y., Shen, D. et al. (2019) Tensor generalized estimating equations for longitudinal imaging analysis. *Statistica Sinica, 29*, 1977–2005.

Zhao, Q., Zhou, G., Zhang, L. and Cichocki, A. (2014) Tensor-variate gaussian processes regression and its application to video surveillance. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1265–1269. IEEE.

Zhou, H. and Li, L. (2014) Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*, 463–483.

Zhou, H., Li, L. and Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association, 108*, 540–552.

Zhou, S., Shen, X., Wolfe, D. et al. (1998) Local asymptotics for regression splines and confidence regions. *The Annals of Statistics, 26*, 1760–1782.

Zhu, Z., Hu, X. and Caverlee, J. (2018) Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1153–1162. ACM.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301–320.
Supplementary Material

A  Asymptotic study

A.1  Notations

We use $C$ with or without subscripts to represent generic constants that may change values from line to line. To simplify the notations, we let

$$J = \{j = (i_1, \cdots, i_D) : 1 \leq i_d \leq p_d, d = 1, \ldots, D\}. \quad (A.1)$$

By noting that

$$s = \prod_{d=1}^{D} p_d,$$

we have the cardinality $|J| = s$. The Hilbert-Schmidt norm of a generic tensor $A$ is defined as $\|A\|_{{HS}} = \langle A, A \rangle^{1/2}$.

The concept of Gaussian width (Chandrasekaran et al., 2012; Vershynin, 2018) and $\gamma$-functionals (Talagrand, 2005; Banerjee et al., 2015) will be used in several places of our proofs. We put their definitions in the beginning of technical results.

**Definition 2** (Gaussian width). For any set $P \subset \mathbb{R}^p$, the Gaussian width of the set $P$ is defined as

$$w(P) = \mathbb{E}_x \sup_{a \in P} \langle a, x \rangle,$$

where the expectation is over $x \sim N(0, I_{p \times p})$, a vector of independently standard Gaussian random variables.

**Definition 3** ($\gamma$-functionals). Consider a metric space $(T, d)$ and for a finite set $A \subset T$, let $|A|$ denote its cardinality. An admissible sequence is an increasing sequence of subsets $\{A_n, n \geq 0\}$ of $T$, such that $|A_0| = 1$ and for $n \geq 1$, $|A_n| = 2^n$. Given $\alpha > 0$, we define the $\gamma_\alpha$-functional as

$$\gamma_\alpha(T, d) = \inf \sup_{t \in T} \sum_{n=0}^{\infty} \text{Diam}\{A_n(t)\},$$

where $A_n(t)$ is the unique element of $A_n$ that contains $t$, $\text{Diam}\{A_n(t)\}$ is the diameter of $A_n$ according to $d$, and the infimum is over all admissible sequences of $T$.

For convenience, we use a mapping $\Omega : \mathbb{R}^{p_1 \times \cdots \times p_D \times K} \times \mathbb{R} \to \mathbb{R}^{p_1 \times \cdots \times p_D \times K}$ to represent the operator of absorbing the constant into the coefficients of B-spline basis for the first predictor. More precisely, $\Omega$ is defined by

$$A^b = \Omega(A, \nu), \quad (A.2)$$

where $A^b_{i_1, \cdots, i_D, k} = A_{i_1, \cdots, i_D, k}$, for $(i_1, \cdots, i_D) \neq (1, \cdots, 1)$ and $A^b_{1, \cdots, 1, k} = A_{1, \cdots, 1, k} + s\nu$, $k = 1, \ldots, K$. It then follows from the property of B-spline functions that

$$\nu + \frac{1}{s} \langle A, \Phi(X) \rangle = \frac{1}{s} \langle A^b, \Phi(X) \rangle.$$

This property simplifies the development of the asymptotic theory since $A^b$ still enjoys a CP structure.
We also write $A_0 = \sum_{r=1}^{R_0} B_{0r} \circ \alpha_{0r}$, $r = 1, \ldots, R$, where $\alpha_{0r}$ satisfies (15) and (16). We define
\[
\tilde{h}_n = \max \left\{ h_n^{1/(-\log h_n)}, h_n \right\},
\]
and this definition, we have
\[
\tilde{h}_n^2 h_n^{-2} \asymp h_n^{-2/(\log h_n)} (\log^{-2} h_n),
\]
which is a quantity presented in the sample size requirements of Theorem 1 and Corollary 1.

A.2 Proof of Theorem 1
Suppose $(\hat{A}_{PLS}, \hat{\nu}_{PLS})$ is a solution to (10) and
\[
\hat{A}_{PLS} = \sum_{r=1}^{R} \hat{\beta}_{r,1} \circ \hat{\beta}_{r,2} \circ \cdots \circ \hat{\beta}_{r,D} \circ \hat{\alpha}_r.
\]
In the following, we let
\[
\hat{G} = G(\hat{\theta}_{PLS}),
\]
where $\hat{\theta}_{PLS} = (\hat{B}_1, \ldots, \hat{B}_D)$, $\hat{B}_d = (\hat{\beta}_{1,d}, \ldots, \hat{\beta}_{R,d})$, $d = 1, \ldots, D$, $\hat{B}_{D+1} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_R)$. By Lemma B.1, there exists $\nu_{PLS} \in \mathbb{R}$ and
\[
\tilde{A}_{PLS} = \sum_{r=1}^{R} \hat{\beta}_{r,1} \circ \hat{\beta}_{r,2} \circ \cdots \circ \hat{\beta}_{r,D} \circ \hat{\alpha}_r \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K},
\]
such that
\[
\hat{\nu}_{PLS} + \frac{1}{s} \langle \hat{A}_{PLS}, \Phi(X) \rangle = \nu_{PLS} + \frac{1}{s} \langle \tilde{A}_{PLS}, \tilde{\Phi}(X) \rangle,
\]
where $\hat{\alpha}_r = (\hat{\alpha}_{r,1}, \ldots, \hat{\alpha}_{r,K})^\top$ satisfying
\[
\sum_{k=1}^{K} \hat{\alpha}_{r,k} u_k = 0,
\]
with $u_k = \int_0^1 b_k(x) dx$. Recall that
\[
A_0 = \sum_{r=1}^{R_0} \beta_{0r,1} \circ \beta_{0r,2} \circ \cdots \circ \beta_{0r,D} \circ \alpha_{0r}, \quad \alpha_{0r} = (\alpha_{0r,1}, \ldots, \alpha_{0r,K})^\top, \quad \sum_{k=1}^{K} \alpha_{0r,k} u_k = 0.
\]
The proof of Lemma B.1 also shows that there exists $\tilde{\nu}_0 \in \mathbb{R}$ and
\[
\tilde{A}_0 = \sum_{r=1}^{R_0} \beta_{0r,1} \circ \beta_{0r,2} \circ \cdots \circ \beta_{0r,D} \circ \alpha_{0r} \in \mathbb{R}^{p_1 \times \cdots \times p_D \times (K-1)},
\]
such that
\[
\tilde{\nu}_0 + \frac{1}{s} \langle \tilde{A}_0, \tilde{\Phi}(X) \rangle = \nu_0 + \frac{1}{s} \langle A_0, \Phi(X) \rangle.
\]
To achieve the restrictions in the optimization problem (10), we normalize \( \tilde{\alpha}_{0r} \) in \( \tilde{A}_0 \) by

\[
\tilde{A}_0 = \sum_{r=1}^{R_0} \left( \| \tilde{\alpha}_{0r} \|_2 \cdot \beta_{0r,1} \right) \circ \beta_{0r,2} \circ \cdots \circ \beta_{0r,D} \circ \tilde{\alpha}_{0r} / \| \tilde{\alpha}_{0r} \|_2.
\]

Using the rescaling strategy (13) on \( \{ \| \tilde{\alpha}_{0r} \|_2 \beta_{0r,1}, \beta_{0r,2}, \ldots, \beta_{0r,D} \} \) for \( r = 1, \ldots, R \), we get a solution \( \{ \rho_{0r,d} \}_{d=1}^{D}, \ r = 1, \ldots, R \). Denoting \( \tilde{\beta}_{0r,1} = \rho_{0r,1} \| \tilde{\alpha}_{0r} \|_2 \beta_{0r,1}, \ \tilde{\beta}_{0r,d} = \rho_{0r,d} \beta_{0r,d}, \ d = 2, \ldots, D, \) we have

\[
\tilde{A}_0 = \sum_{r=1}^{R_0} \tilde{\beta}_{0r,1} \circ \tilde{\beta}_{0r,2} \circ \cdots \circ \tilde{\beta}_{0r,D} \circ \tilde{\alpha}_{0r} / \| \tilde{\alpha}_{0r} \|_2.
\]

Let

\[
G_0 = G(\tilde{\theta}_0),
\]

where \( \tilde{\theta}_0 = (\tilde{\nu}_0, \tilde{B}_{0,1}, \tilde{B}_{0,D}, \tilde{B}_{0,D+1}), \ \tilde{B}_{0,d} = (\tilde{\beta}_{01,d}, \ldots, \tilde{\beta}_{0R,d}), \ d = 1, \ldots, D, \) and

\[
\tilde{B}_{0,D+1} = \left( \frac{\tilde{\alpha}_{01}}{\| \tilde{\alpha}_{01} \|_2}, \ldots, \frac{\tilde{\alpha}_{0R}}{\| \tilde{\alpha}_{0R} \|_2} \right).
\]

Using (A.6)–(A.9), we obtain

\[
\sum_{i=1}^{n} \left( y_i - \tilde{\nu}_{PLS} - \frac{1}{s} \langle A_{PLS}, \Phi(X_i) \rangle \right)^2 + \tilde{G} \leq \sum_{i=1}^{n} \left( y_i - \nu_0 - \frac{1}{s} \langle A_0, \Phi(X_i) \rangle \right)^2 + G_0. \quad \text{(A.10)}
\]

Let \( \tilde{A}_{PLS}^b = \Omega(\tilde{A}_{PLS}, \tilde{\nu}_{PLS}) \) and \( A_0^b = \Omega(A_0, \nu_0) \). Since \( \tilde{G} \geq 0 \), it implies that

\[
\sum_{i=1}^{n} \left( y_i - \frac{1}{s} \langle A_{PLS}^b, \Phi(X_i) \rangle \right)^2 \leq \sum_{i=1}^{n} \left( y_i - \frac{1}{s} \langle A_0^b, \Phi(X_i) \rangle \right)^2 + G_0. \quad \text{(A.11)}
\]

Let \( A_{PLS}^z = \tilde{A}_{PLS} - A_0^z, \ a_{PLS}^z = \text{vec}(A_{PLS}^z), \ a_0^z = \text{vec}(A_0^z) \) and

\[
Z = (z_1, \ldots, z_n)^T \in \mathbb{R}^{n \times sK}. \quad \text{(A.12)}
\]

where \( z_i = \text{vec}\{\Phi(X_i)\}, \ i = 1, \ldots, n \). Using (A.11) and working out the squares, we obtain

\[
\frac{1}{s^2} \| Z_{PLS}^z \|_2^2 \leq 2 \left( \frac{1}{s} \| Z_{PLS}^z \|, \epsilon \right) + 2 \left( \frac{1}{s} Z_{PLS}^z, y - \epsilon - \frac{1}{s} Z_0^z \right) + G_0. \quad \text{(A.13)}
\]

where \( y = (y_1, \ldots, y_n)^T \). We will finish the proof by taking the union of probabilities of the events (A.23) and (A.27) in Sections A.2.1 and A.2.2, respectively.

**A.2.1 Bound of CP parameters without restriction**

In the subsection, we show the upper bound of the CP parameters without the scale restriction induced by penalization. By (A.7), Lemmas A.1 and A.6, we have \( \sum_{k=1}^{K} A_{j,k} u_k = 0 \) for \( j \in J/\{1, \cdots, 1\} \). Since \( \text{rank}(A_0^z) \leq R_0 + 1, \ \text{rank}(\tilde{A}_{PLS}^z) \leq R + 1 \), it is trivial to see
\( \text{rank}(\mathbf{A}_\text{PLS}^\sharp) \leq R_0 + R + 2 \). To finish the proof, we will try to find the upper bound of the right hand side and the lower bound of the left hand side with respect to \( \| \mathbf{a}_\text{PLS}^\sharp \|_2 \) in (A.13).

First, we will find the upper bound of \( \langle \mathbf{Za}_\text{PLS}^\sharp, \epsilon \rangle \). To simplify the notations, let

\[
\mathcal{P}_1 = \left\{ \frac{\text{vec} (\mathbf{A})}{\| \mathbf{A} \|_{HS}} : \sum_{k=1}^{K} A_{j,k} u_k = 0, \text{ for } j \in J/\{(1, \ldots, 1)\}, \text{rank}(\mathbf{A}) \leq R_1 \right\},
\]

where \( R_1 = R + R_0 + 2 \). Since \( R_0 \leq R \), we have

\[
R_1 \leq 2R + 2.
\]

By Lemma A.4, to show

\[
C_1 nh_n \| \mathbf{a}_\text{PLS}^\sharp \|_2^2 \leq \| \mathbf{Za}_\text{PLS}^\sharp \|_2^2 \leq C_2 nh_n \| \mathbf{a}_\text{PLS}^\sharp \|_2^2,
\]

with probability at least \( 1 - 2 \exp \{- C_3 w^2(\mathcal{P}_1) \} \), we only need to prove \( n > C \tilde{h}_n^2 h_n^{-2} w^2(\mathcal{P}_1) \).

By (A.15) and Lemma A.3, the Gaussian width

\[
w(\mathcal{P}_1) \leq C_4 \left( R_1^{D+1} + R_1 \sum_{d=1}^{D} p_d + R_1 K \right)^{1/2} \leq C_5 \left( R_1^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}.
\]

Due to the assumption that \( n > C_6 \tilde{h}_n^2 h_n^{-2} (R_1^{D+1} + R \sum_{i=1}^{D} p_i + RK) \), (A.16) is thus satisfied with probability at least

\[
1 - 2 \exp \left\{ - C \left( R_1^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.
\]

By (A.14)–(A.17) and Lemma A.5, we have the following upper bound

\[
\langle \mathbf{Za}_\text{PLS}^\sharp, \epsilon \rangle \leq C_7 \| \mathbf{a}_\text{PLS}^\sharp \|_2 \left\{ nh_n \left( R_1^{D+1} + \sum_{i=1}^{D} R p_i + RK \right) \right\}^{1/2},
\]

with probability at least

\[
1 - C_8 \exp \left\{ - C_9 \left( R_1^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.
\]

Second, we find the upper bound of \( \langle \mathbf{Za}_\text{PLS}^\sharp, \mathbf{y} - \mathbf{Za}_0^\flat \rangle \). Note that

\[
\left\| \mathbf{y} - \epsilon - \frac{1}{s} \mathbf{Za}_0^\flat \right\|_2^2 = \sum_{i=1}^{n} \frac{1}{s} \sum_{r=1}^{R_0} \left\{ \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{A}_0, \Phi(\mathbf{X}_i) \rangle \right\}^2 \\
\leq \sum_{i=1}^{n} \left\{ \frac{1}{s} \sum_{r=1}^{R_0} \left\{ \langle \mathbf{B}_{0r}, F_r(\mathbf{X}_i) \rangle - \langle \mathbf{B}_{0r} \circ \mathbf{a}_{0r}, \Phi(\mathbf{X}_i) \rangle \right\}^2 \\
\leq \sum_{i=1}^{n} \left\{ \frac{1}{s} \sum_{r=1}^{R_0} C_1 \| \text{vec}(\mathbf{B}_{0r}) \|_1 \right\}^2 \\
= C_2 \left\{ \frac{\sum_{r=1}^{R_0} \| \text{vec}(\mathbf{B}_{0r}) \|_1}{s} \right\}^2 \frac{n}{K^{2\tau}}.
\]
Using the Cauchy-Schwarz inequality, (A.16), (A.17) and (A.19), it shows that
\[
\left\langle \frac{1}{s} Z_{\text{PLS}}^\sharp \mathbf{y}, \mathbf{y} - \frac{1}{s} \mathbf{Za}_0 \right\rangle \leq \left\| \mathbf{y} - \frac{1}{s} \mathbf{Za}_0 \right\|_2 \left\| \frac{1}{s} Z_{\text{PLS}}^\sharp \right\|_2 \\
\leq \frac{C_3}{s} \| \mathbf{Za}_{\text{PLS}} \|_2 \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(\mathbf{B}_0r) \|_1}{s} \right\} \frac{\sqrt{n}}{K^\tau} \\
\leq \frac{C_4}{s} \| \mathbf{a}_{\text{PLS}} \|_2 \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(\mathbf{B}_0r) \|_1}{s} \right\} \frac{\sqrt{n}}{K^\tau} 
\right)
\] 
(A.20)

with probability at least
\[
1 - C_5 \exp \left\{ - C_6 \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.
\]

Third, applying (A.16), (A.18) and (A.20) to (A.13), we get
\[
\frac{C_7}{s^2} \| \mathbf{a}_{\text{PLS}} \|^2 \leq \frac{\delta_3}{s} \| \mathbf{a}_{\text{PLS}} \|^2 + \frac{1}{nh_n} G_0,
\]
(A.21)

with probability at least
\[
1 - C_1 \exp \left\{ - C_2 \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\},
\]
(A.22)

where
\[
\delta_3 = C_3 \left\{ \frac{K (R^{D+1} + \sum_{i=1}^{D} R p_i + RK)}{n} \right\}^{1/2} + C_4 \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(\mathbf{B}_0r) \|_1}{s} \right\} \frac{1}{K^\tau - 1/2}.
\]

By solving the second order inequality (A.21), we obtain
\[
\frac{C_5}{s} \| \mathbf{a}_{\text{PLS}} \|^2 \leq \frac{\left\{ \delta_3^2 + 4G_0/(nh_n) \right\}^{1/2} + \delta_3}{2},
\]

with the probability at least (A.22). Further, by Assumption 2 and (A.38) of Lemma A.2, we have
\[
\| \hat{m}_{\text{PLS}} - m_0 \|^2 \leq C_6 h_n \frac{1}{s^2} \| \hat{A}_{\text{PLS}} - A_0 \|_H^2 = C_6 h_n \frac{1}{s^2} \| \mathbf{a}_{\text{PLS}} \|^2.
\]

The displayed two equations in the above together with (A.4) imply that
\[
\| \hat{m}_{\text{PLS}} - m_0 \|^2 \leq \frac{C_7 \{ \delta_3^2 + (4KG_0)/n \}}{K},
\]
(A.23)

with the probability at least (A.22), which completes the proof of one case.
A.2.2 Bound of CP parameters with restriction

In the subsection, we show the upper bound of the CP parameters under the scale restriction induced by penalization. Let

\[
\delta_B = \frac{1}{\varpi(n)} \left[ \left\{ \sum_{r=1}^{R_0} \|\text{vec}(B_{0r})\|_1 \right\} \frac{C_4}{K^{2\tau}} + C_5 n + G_0 \right]^{S_G}
\]

and

\[
\delta_v = \frac{C_1}{s} \sum_{r=1}^{R_0} \|\text{vec}(B_{0r})\|_1 + |\nu_0| + C_2 + C_3 \sqrt{K R \delta_B^{D/2}} / \sqrt{s}.
\]

By Lemma A.9, we have

\[
\sum_{d=1}^{D} \sum_{r=1}^{R} \|\hat{\beta}_{r,d}\|_2^2 \leq \delta_B,
\]

and

\[
|\hat{\nu}| \leq \delta_v,
\]

with probability at least \(1 - C_6 \exp(-C_7 n)\). It then shows that

\[
A_{\text{PLS}}^* = A_{\text{PLS}}^* \mathbf{1}_{\left\{ \sum_{d=1}^{D} \sum_{r=1}^{R} \|\hat{\beta}_{r,d}\|_2^2 \leq \delta_B, |\hat{\nu}| \leq \delta_v \right\}}, \tag{A.24}
\]

with probability at least \(1 - C_6 \exp(-C_7 n)\), where \(\mathbf{1}_{\cdot}\) is an indicator function. The assumption of sample size \(n\) implies that

\[
n > C_9 \tilde{h}_n^{-2} \left( R^2 \log \delta_{\text{pen}} + R \sum_{d=1}^{D+1} p_d \right), \tag{A.25}
\]

where

\[
\delta_{\text{pen}} = n \max \left\{ C_5 \left( \delta_v s + \sqrt{s} R \delta_B^{D/2} \right)^{2/D} , C_6 R K^2 \sum_{r=1}^{R_0} \|\theta_{0r,d}\|_2^2 + (s \nu_0)^{2/D} , \sum_{r=1}^{R_0} \|\alpha_{0r}\|_2^2 + 1 \right\}. \tag{A.26}
\]

Using the arguments as in the proof of Lemma A.11 and applying (A.24) to (A.13), we thus obtain

\[
\|\hat{m}_{\text{PLS}} - m_0\|^2 \leq C_9 \frac{R^2 \log \delta_{\text{pen}} + R \sum_{d=1}^{D} p_d + R K}{n} + C_1 \left\{ \sum_{r=1}^{R_0} \|\text{vec}(B_{0r})\|_1 \right\} \frac{1}{K^{2\tau}} + \frac{C_2}{n} G_0, \tag{A.27}
\]

with probability at least

\[
1 - C_3 \exp \left\{ - C_4 \left( R^2 \log \delta_{\text{pen}} + R \sum_{d=1}^{D} p_d + R K \right) \right\}.
\]

A.3 Proof of Corollary 1

Proof. It is a special case of Theorem 1 where the bias term \(G_0 = 0\) due to elimination of the penalty function. Using the proof in Section A.2.1 under Assumptions 2–5, we obtain the result of (19). \(\square\)
A.4 Proof of Corollary 2

Proof. When we take the elastic-net penalty form (14) in (12) with \( \lambda_2 = 0 \), by Lemma A.10, the penalty function satisfies Assumption 1 with \( \varpi(n) = C\lambda_1 \) and \( S_G = 1 \). By Lemma A.15,

\[
\delta_B \leq \frac{C}{\lambda_1} \left[ \left\{ \sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1 \right\}^2 \frac{1}{K^{2\tau}} + n + G_0 \right] \\
\leq \frac{1}{\lambda_1} \left\{ C_1 R_0^2 + C_2 n + C_3 \lambda_1 R_0(K/C_4)^{2K/D} \sum_{d=1}^{D} p_d^2 \right\} \\
\leq \frac{C_1 R_0^2 + C_2 n}{\lambda_1} + C_3 R_0(K/C_4)^{2K/D} \sum_{d=1}^{D} p_d^2.
\]

Taking

\[
\frac{C_5}{R_0(K/C_4)^{2K/D} \sum_{d=1}^{D} p_d^2} \leq \lambda_1 \leq \frac{C_6 R^2 + \sum d p_d R + KR}{C_3 R_0(K/C_4)^{2K/D} \sum_{d=1}^{D} p_d^2}
\]  (A.28)

yields

\[
G_0 \leq C \left( R^2 + \sum d p_d R + KR \right)
\]

and

\[
\delta_B \leq C_1 n C_2 (K/C_3)^{C_4 K}.
\]

It implies \( \delta_{\text{pen}} \leq C_1 n C_2 (K/C_3)^{C_4 K} \) and thus

\[
\log \delta_{\text{pen}} \leq C_2 \log n + C_4 K \log(K/C_3) \leq C_5 K \log n,
\]

which completes the proof. \( \square \)

A.5 Proof of Theorem 2

Proof. We let \( R \geq 4 \) and \( \gamma \) be a positive constant for simplicity. By Lemmas A.12 and A.13 together with the proof of Theorem 4 in Suzuki (2015), we can find a set \( \mathcal{A}_M \subset \mathbb{R}^{p_1 \times \cdots \times p_D \times M} \) satisfying

\[
|\mathcal{A}_M| \geq C_1^R \prod_{d=1}^{D} \left( \frac{1}{\varrho} \right)^{C_2 R p_d - C_3 R + C_4 M R},
\]  (A.29)

and

\[
\|\mathbf{A}_M - \mathbf{A}'_M\|^2 \geq C_5 (\gamma^2)^{D} \varrho^2 M,
\]  (A.30)

where \( \mathbf{A}_M, \mathbf{A}'_M \in \mathcal{A}_M, \mathbf{A}_M \neq \mathbf{A}'_M, \) and \( \max_d p_d \geq R \). To more specific, by Lemma A.13, there exists a subset \( \mathcal{W} \subset \{ \mathbf{W} \in \mathbb{R}^{M \times R}, \| \mathbf{W}_{:,r} \|_2 = M/R \} \) such that the cardinality

\[
|\mathcal{W}| \geq C_1^R \left( \frac{1}{\varrho} \right)^{MR/4 - 2R},
\]

\[
\|\mathbf{W} - \mathbf{W}'\|_2 \geq C_2 \varrho^2 M R,
\]
and 
\[ \|W_{:,r} - W'_{:,r}\|^2 \leq M, \]
where \(0 < C_1, C_2 < 1, 0 < \varrho \leq C < 1, W, W' \in \mathcal{W}\) and \(W \neq W'\). Suppose \(d^* \in \arg\max_d p_d\) and \(p_{d^*}/R\) is an integer. The requirement of \(p_{d^*}/R\) to be an integer is for the notational simplicity. We only need \(p_{d^*} \geq R\). By Lemma A.12, there exists a set \(B \subset \{B \in \mathbb{R}^{(p_{d^*}/R) \times R}, \|B_{:,r}\|^2 = \gamma^2/R\}\) satisfying that
\[ |B| \geq C R \left( \frac{1}{\varrho} \right)^{C_2 R p_{d^*} - C_3 R}, \]
\[ \|B - B'\|^2 \geq C_4 \gamma^2 \varrho^2, \]
\[ \|B_{:,r} - B'_{:,r}\|_2 \leq \gamma/R, \]
where \(B' \in B\) and \(B \neq B'\). We construct \(B_{d^*} \subset \mathbb{R}^{p_{d^*} \times R}\) as follow
\[ B_{d^*} = \left\{ B = \sqrt{R} \begin{pmatrix} b_1 & 0 & \cdots & 0 \\ 0 & b_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & b_R \end{pmatrix} \left( \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_R \end{pmatrix} \right) \in B \right\}. \]
We denote \(B_d, d = 1, \ldots, d^* - 1, d^* + 1, \ldots, D\) as the set in Lemma A.12 and define \(A_M\) as
\[ A_M = \left\{ A_M : A_M = \sum_{r=1}^{R} \beta_{r,1} \circ \cdots \circ \beta_{r,D} \circ w_r, (\beta_{1,d}, \ldots, \beta_{R,d}) \in B_d, (w_1, \ldots, w_R) \in \mathcal{W} \right\}, \]
which yields (A.29). Suppose \(A_M, A'_M \in A_M\) and \(A_M \neq A'_M\), we then have
\[ \|A_M - A'_M\|^2_{HS} = \left\| \sum_{r=1}^{R} (\beta_{r,1} \circ \cdots \circ \beta_{r,D} \circ w_r - \beta'_{r,1} \circ \cdots \circ \beta'_{r,D} \circ w'_r) \right\|^2 \]
\[ = \sum_{r=1}^{R} \left\| \beta_{r,1} \circ \cdots \circ \beta_{r,D} \circ w_r - \beta'_{r,1} \circ \cdots \circ \beta'_{r,D} \circ w'_r \right\|^2. \]
Due to \(A_M \neq A'_M\), there must at least one component be different. There are two possible cases:

Case 1: For some \(d\), \((\beta_{1,d}, \ldots, \beta_{R,d})\) and \((\beta'_{1,d}, \ldots, \beta'_{R,d})\) are different.

Case 2: \((w_1, \ldots, w_R)\) and \((w'_1, \ldots, w'_R)\) are different.

Our following analysis is based upon these two cases.

For Case 1, by (A.31), we have
\[ \|(\beta_{1,d}, \ldots, \beta_{R,d}) - (\beta'_{1,d}, \ldots, \beta'_{R,d})\|^2 \geq C_2 \gamma^2 \varrho^2 R, \]
which yields that there exists a \(r\) satisfying
\[ \|\beta_{r,d} - \beta'_{r,d}\|^2 \geq C_2 \gamma^2 \varrho^2. \]
By the inequality (26) in Suzuki (2015) and its related equation, we have
\[
\|\beta_{r,1} \circ \cdots \circ \beta_{r,D} \circ w_r - \beta'_{r,1} \circ \cdots \circ \beta'_{r,D} \circ w'_r\|^2 \\
= \|\beta_{r,d} \circ \beta_{r,1} \cdots \circ \beta_{r,d-1} \circ \beta_{r,d+1} \cdots \circ \beta_{r,D} \circ w_r \\
- \beta'_{r,d} \circ \beta'_{r,1} \cdots \circ \beta'_{r,d-1} \circ \beta'_{r,d+1} \cdots \circ \beta'_{r,D} \circ w'_r\|^2 \\
= \|\beta_{r,d} - \beta'_{r,d}\|^2 \|U_{-d}\|^2 + \|U_{-d} - U'_{-d}\|^2 \left(\|\beta_{r,d}\|^2 - \frac{1}{2}\|\beta_{r,d} - \beta'_{r,d}\|^2\right) \\
\geq \|\beta_{r,d} - \beta'_{r,d}\|^2 \prod_{d' \neq d} \|\beta_{r,d}\|^2 \|w_r\|^2 \\
\geq C \varrho^2 (\gamma^2) D M,
\]
(A.33)
where \(U_{-d} = \beta_{r,1} \cdots \beta_{r,d-1} \circ \beta_{r,d+1} \cdots \beta_{r,D} \circ w_r\) and \(U'_{-d}\) is defined analogously. For Case 2, similar to Case 1, there exists a \(r\) satisfying \(|w_r - w'_r|^2 \geq C \varrho^2 M\), and we have
\[
\|\beta_{r,1} \circ \cdots \circ \beta_{r,D} \circ w_r - \beta'_{r,1} \circ \cdots \circ \beta'_{r,D} \circ w'_r\|^2 \geq C \varrho^2 (\gamma^2) D M.
\]
(A.34)

Using (A.32)–(A.34) yields (A.30).

Define \(\phi_k(x)\) as (6.5) of Yang and Tokdar (2015). Denote
\[
\mathcal{F}_M = \left\{ f_w(x) = \sum_{k=1}^{M} w_k \phi_k(x), w = (w_1, \ldots, w_M)^\top \in \mathbb{R}^M \right\},
\]
where \(M\) is the minimal integer larger than \(1/(2\tilde{h})\) and \(\tilde{h} \in (0, 1/2)\). For any \(f_w \in \mathcal{F}_M\), we have
\[
\int_0^1 f_w^2(x) dx = \|w\|_2^2 \tilde{h}^{2r+1}
\]
and
\[
\int_0^1 f_w(x) dx = 0.
\]

By definition, for any \(B_r^{(1)}, B_r^{(2)} \in \mathbb{R}^{p \times \cdots \times pD}\) and \(f_{w_1,r}, f_{w_2,r} \in \mathcal{F}_M\), we have
\[
\left\| \left\langle \sum_{r=1}^{R} B_r^{(1)}, F_{w_1,r}(X) \right\rangle - \left\langle \sum_{r=1}^{R} B_r^{(2)}, F_{w_2,r}(X) \right\rangle \right\|^2
\]
\[
= \left\| \left\langle \sum_{r=1}^{R} B_r^{(1)} \circ \Phi_{w}(X) \right\rangle - \left\langle \sum_{r=1}^{R} B_r^{(2)} \circ \Phi_{w}(X) \right\rangle \right\|^2 \\
= C \left\| \sum_{r=1}^{R} B_r^{(1)} \circ \Phi_{w}(X) - \sum_{r=1}^{R} B_r^{(2)} \circ \Phi_{w}(X) \right\|^2 \tilde{h}^{2r+1},
\]
where \(\{\Phi_{w}(X)\}_{i_1, \ldots, i_D, k} = \phi_k(X_{i_1, \ldots, i_D})\) and \(\{F_{w,l,r}(X)\}_{i_1, \ldots, i_D, k} = f_{w,l}(X_{i_1, \ldots, i_D}), l = 1, 2\).

Denote
\[
\mathcal{M}_{03} = \left\{ m(X) : m(X) = \sum_{r=1}^{R} \left\langle \tilde{\beta}_{r,1} \circ \cdots \circ \tilde{\beta}_{r,D} \circ w_r, \Phi_{w}(X) \right\rangle, \right. \\
\left. \sum_{r=1}^{R} \tilde{\beta}_{r,1} \circ \cdots \circ \tilde{\beta}_{r,D} \circ w_r \in A_M \right\}.
\]
(A.35)
Thus, the above discussions lead to

$$|\mathcal{M}_{03}| \geq C_1^R \prod_{d=1}^D \left( \frac{1}{\varrho/\gamma^D} \right)^{C_2 R p_d - C_3 R + C_4 R/\bar{h}}$$

and

$$\|m - m'\|^2 \geq C_5 \varrho^2 \bar{h}^{2\tau},$$

where $m, m' \in \mathcal{M}_{03}$ and $m \neq m'$. In order to use Theorem 1 in Yang and Barron (1999), we only need to find the order of $(\varrho \bar{h}^{\tau})^2$ such that

$$C_6 \left( \frac{1}{\varrho/\gamma^D} \right)^{C_2 R p_d - C_3 R + C_4 R/\bar{h}} \leq \log \left\{ \min \left\{ \sum_{d=1}^D R_{p_d} \rho^2 \gamma^2 D, \rho^2 \gamma^2 D \right\} \right\} \quad (A.36)$$

is satisfied. Since $\max_{d} p_d$ is bounded from below, it can be shown that $\bar{h} \leq C_5$ together with

$$\varrho^2 = C_6 \min \left\{ \sum_{d=1}^D R_{p_d} \rho^2 \gamma^2 D, \rho^2 \gamma^2 D \right\},$$

is one sufficient condition of (A.36). This case implies that

$$(\varrho \bar{h}^{\tau})^2 = C \min \left\{ \sum_{d=1}^D R_{p_d} \rho^2 \gamma^2 D, \rho^2 \gamma^2 D \right\}.$$  

On the other hand, $\varrho = C_7 \gamma^D$ and

$$\bar{h}^{2\tau} = C_8 \left( \frac{R}{n \varrho^2} \right)^{2\tau/\tau + 1},$$

will also yield (A.36), which implies

$$(\varrho \bar{h}^{\tau})^2 = C_9 \left( \frac{R}{n \varrho^2} \right)^{2\tau/\tau + 1} = C_2 \gamma \left( \frac{R}{n} \right)^{2\tau/\tau + 1}. $$

Using the proofs of Theorem 1 in Yang and Barron (1999) and Theorem 4 in Suzuki (2015), we obtain

$$\inf \sup_{m \in \mathcal{M}_{03}} \mathbb{E} \|m - \hat{m}\|^2 \geq \max \left\{ \min \left\{ \sum_{d=1}^D R_{p_d} \rho^2 \gamma^2 D, \rho^2 \gamma^2 D \right\}, \gamma \left( \frac{R}{n} \right)^{2\tau/\tau + 1} \right\}. \quad (A.37)$$

Finally, recall

$$\mathcal{M}_{00} = \left\{ m(\mathbf{X}) : m(\mathbf{X}) = \nu + \sum_{r=1}^R \left( \frac{\beta_{r,1}}{p_1} \circ \ldots \circ \frac{\beta_{r,D}}{p_D}, F_r(\mathbf{X}) \right), \nu \leq C_1, \left\| \frac{\beta_{r,d}}{p_d} \right\|_{1} \leq \gamma, \beta_{r,d} \in \mathbb{R}^{p_d}, \int_0^1 f_r^2(t) dt \leq C_2 \right\}.$$
Let
\[ M_{01} = \left\{ m(X) : m(X) = \nu + \sum_{r=1}^{R} \left\langle \frac{\beta_{r,1}}{p_1} \circ \ldots \circ \frac{\beta_{r,D}}{p_D}, F_r(X) \right\rangle, \right. \]
\[ \left. \quad \nu \leq C_1, \left\| \frac{\beta_{r,d}}{p_d} \right\|_2^2 \leq \gamma^2, \beta_{r,d} \in \mathbb{R}^{p_d}, \int_0^1 f_r^2(t) \, dt \leq C_2 \right\}, \]
and
\[ M_{02} = \left\{ m(X) : m(X) = \sum_{r=1}^{R} \left\langle \tilde{\beta}_{r,1} \circ \ldots \circ \tilde{\beta}_{r,D}, F_r(X) \right\rangle, \right. \]
\[ \left. \quad ||\tilde{\beta}_{r,d}||_2^2 \leq \gamma^2, \tilde{\beta}_{r,d} \in \mathbb{R}^{p_d}, \int_0^1 f_r^2(t) \, dt \leq C_2 \right\}. \]
Due to the definition of \( M_{03} \) in (A.35), it is each to check that
\[ M_{03} \subset M_{02} \subset M_{01} \subset M_{00}. \]
Thus,
\[ \inf_{m \in M_{00}} \sup \mathbb{E} \| m - \hat{m} \|^2 \geq \inf_{m \in M_{03}} \sup \mathbb{E} \| m - \hat{m} \|^2, \]
which finish the proof. If \( R \leq 3 \), we use Lemma A.14 to replace the arguments about (A.85) in the proof of Lemma A.12 and can obtain a representation like (A.37).

### A.6 Technical results

**Lemma A.1.** Suppose \( A \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K} \) has such a CP decomposition,
\[ A = \sum_{r=1}^{R} \beta_{r,1} \circ \ldots \circ \beta_{r,D} \circ \alpha_r, \]
where \( \alpha_r = (\alpha_{r,1}, \ldots, \alpha_{r,K})^T \in \mathbb{R}^K \) and \( \beta_{r,d} \in \mathbb{R}^{p_d} \) for \( d = 1, \ldots, D \) and \( r = 1, \ldots, R \). If
\[ u \in \{(u_1, \ldots, u_K)^T : \sum_{k=1}^{K} \alpha_{r,k} u_k = 0, r = 1, \ldots, R\}, \]
then
\[ \sum_{k=1}^{K} A_{j,k} u_k = 0, \quad \text{for} \quad j \in \mathcal{J}, \]
where \( \mathcal{J} \) is defined in (A.1).

**Proof.** This proof is straightforward. For simplicity, for \( r = 1, \ldots, R \), let
\[ B_r = \beta_{r,1} \circ \ldots \circ \beta_{r,D}. \]
Since
\[ \sum_{k=1}^{K} \alpha_{r,k} u_k = 0, \]
we have
\[ \sum_{k=1}^{K} B_{r,j} \alpha_{r,k} u_k = 0, \quad j \in \mathcal{J}, \]
where $B_{r,j}$ is the $j$-th entry of $B_r$, $r = 1, \ldots, R$. Therefore,
\[
\sum_{k=1}^{K} A_{j,k} u_k = \sum_{k=1}^{K} \sum_{r=1}^{R} B_{r,j} \alpha_{r,k} u_k = \sum_{r=1}^{R} \sum_{k=1}^{K} B_{r,j} \alpha_{r,k} u_k = 0, \quad j \in J.
\]

\[\square\]

Lemma A.2. Suppose $A \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K}$ and $U \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is a random tensor with its entry $U_j \overset{i.i.d.}{\sim} U(0,1)$, for $j \in J$, where $J$ is defined in (A.1). Recall that $(\Phi(X))_{j,k} = b_k(X_j)$, where $\{b_k(x)\}_{k=1}^{K}$ be a B-spline basis, $x \in [0,1]$. Under Assumptions 2 and 5, if $\sum_{k=1}^{K} A_{j,k} u_k = 0$ for $j \in J_1 := J \setminus \{(1, \ldots, 1)\}$, where $u_k = \int_0^1 b_k(x) dx$, then we have

\[i.\]
\[C_1 C_\zeta h_n \|A\|_{HS}^2 \leq \mathbb{E}\{\langle A, \Phi(X) \rangle^2\} \leq C_2 h_n \|A\|_{HS}^2, \tag{A.38}\]

and

\[ii.\]
\[\|\langle A, \Phi(X) \rangle\|_{q_2}^2 \leq C_3 \tilde{h}_n \|A\|_{HS}^2, \tag{A.39}\]

where $C_1, C_2, C_3, C_\zeta$ are positive constants, $C_\zeta$ depends on the order of B-spline $\zeta$, and $\tilde{h}_n$ is defined in (A.3).

Proof. We will prove the population bound (A.38) at first. Let $A_j = (A_{j,1}, \ldots, A_{j,K})^\top$ for $j \in J$. By the property of B-spline (see, e.g., de Boor, 1973, 1976) and Assumption 5, for $1 \leq q \leq +\infty$,
\[C_\zeta \|A_j\|_q \leq h_n^{-1/q} \left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_q \leq C \|A_j\|_q, \tag{A.40}\]

where $C_\zeta$ and $C$ are two positive constants and $C_\zeta$ depends on the order of B-spline $\zeta$. By the independence and the mean zero restriction for $j \in J_1$, we have
\[
\mathbb{E}\{\langle A, \Phi(U) \rangle^2\} = \sum_{j \in J} \mathbb{E} \left[ \left\{ \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\}^2 \right].
\]

Taking $q = 2$ in (A.40) yields
\[C_\zeta h_n \|A_j\|_2 \leq \mathbb{E} \left[ \left\{ \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\}^2 \right] \leq Ch_n \|A_j\|_2^2,
\]

then
\[C_\zeta h_n \|A\|_{HS}^2 \leq \mathbb{E}\{\langle A, \Phi(U) \rangle^2\} \leq Ch_n \|A\|_{HS}^2. \tag{A.41}\]

By Assumption 2, we have
\[C_1 \mathbb{E}\{\langle A, \Phi(U) \rangle^2\} \leq \mathbb{E}\{\langle A, \Phi(X) \rangle^2\} \leq C_4 \mathbb{E}\{\langle A, \Phi(U) \rangle^2\}. \tag{A.42}\]

It follows from (A.41) and (A.42) that
\[C_1 C_\zeta h_n \|A\|_{HS}^2 \leq \mathbb{E}\{\langle A, \Phi(X) \rangle^2\} \leq C_2 h_n \|A\|_{HS}^2,
\]
which completes the proof of (A.38).

Now, we will prove the sub-Gaussian norm bound (A.39). Note that

$$\|\langle \mathbf{A}, \Phi(U) \rangle \|_{\psi_2} \leq \left\| \sum_{j \in J_1} \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2} + \left\| \sum_{k=1}^{K} A_{1,\ldots,1,k} b_k(U_{1,\ldots,1}) \right\|_{\psi_2},$$

then

$$\|\langle \mathbf{A}, \Phi(U) \rangle \|_{\psi_2}^2 \leq 2 \left\| \sum_{j \in J_1} \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2}^2 + 2 \left\| \sum_{k=1}^{K} A_{1,\ldots,1,k} b_k(U_{1,\ldots,1}) \right\|_{\psi_2}^2. \quad (A.43)$$

Using the independence property of $\mathbf{U}$, mean zero restriction of $\mathbf{A}$ and Proposition 2.6.1 of Vershynin (2018), we obtain

$$\left\| \sum_{j \in J_1} \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2}^2 \leq C_5 \sum_{j \in J_1} \left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2}^2. \quad (A.44)$$

It follows from (A.43) and (A.44) that

$$\|\langle \mathbf{A}, \Phi(U) \rangle \|_{\psi_2}^2 \leq 2C_5 \sum_{j \in J_1} \left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2}^2 + 2 \left\| \sum_{k=1}^{K} A_{1,\ldots,1,k} b_k(U_{1,\ldots,1}) \right\|_{\psi_2}^2.$$

Therefore,

$$\|\langle \mathbf{A}, \Phi(U) \rangle \|_{\psi_2}^2 \leq (2C_5 + 2) \sum_{j \in J} \left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{\psi_2}^2 + (2C_5 + 2) \left\| \sum_{k=1}^{K} A_{1,\ldots,1,k} b_k(U_{1,\ldots,1}) \right\|_{\psi_2}^2. \quad (A.45)$$

We then consider the sub-Gaussian norm of $\sum_{k=1}^{K} A_{j,k} b_k(U_j)$. When $q = 1$, by (A.40), we have

$$\left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{1} \leq 2 \left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{2} \leq C \sqrt{h_n} \|\mathbf{A}\|_2. \quad (A.46)$$

Similarly, when $q \geq 2$, we obtain

$$\left\| \sum_{k=1}^{K} A_{j,k} b_k(U_j) \right\|_{q} \leq C \frac{h_n^{1/q}}{\sqrt{q}} \|\mathbf{A}\|_q \leq C \frac{h_n^{1/q}}{\sqrt{q}} \|\mathbf{A}\|_2. \quad (A.47)$$

Since $f(x) = h_n^{1/x} / \sqrt{x}$ get the maximum at $x = -2 \log h_n$, then

$$\frac{h_n^{1/q}}{\sqrt{q}} \|\mathbf{A}\|_2 \leq \frac{h_n^{1/(-2 \log h_n)}}{(-2 \log h_n)^{1/2}} \|\mathbf{A}\|_2. \quad (A.48)$$
Due to the definition of $\tilde{h}_n$ in (A.3) and using (A.45)–(A.48), we get

$$\|\langle A, \Phi(U) \rangle \|_{\psi_2}^2 \leq (2C_5 + 2)\tilde{h}_n C^2 \|A\|_{HS}^2. \tag{A.49}$$

Note that for $q \geq 1$,

$$\frac{1}{\sqrt{q}} \left[ \mathbb{E}\{ \langle A, \Phi(X) \rangle^q \} \right]^{1/q} \leq C \frac{1}{\sqrt{q}} \left[ \mathbb{E}\{ \| A, \Phi(U) \|^q \} \right]^{1/q} \leq C \|\langle A, \Phi(U) \rangle \|_{\psi_2},$$

therefore,

$$\|\langle A, \Phi(X) \rangle \|_{\psi_2}^2 \leq C_3 \tilde{h}_n \|A\|_{HS}^2,$$

which completes the proof of (A.39). \qed

**Lemma A.3.** Let $A \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$, and

$$\mathcal{P} = \left\{ \frac{\text{vec}(A)}{\|A\|_{HS}} : \sum_{k=1}^{K} A_{j,k} u_k = 0, \text{ for } j \in J/\{(1, \ldots, 1)\}, \text{ rank}(A) \leq R \right\}, \tag{A.50}$$

where $u_k = \int_0^1 b_k(x)dx$ and $J$ is defined in (A.1). The Gaussian width satisfying

$$w(\mathcal{P}) \leq C \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}. \tag{A.51}$$

**Proof.** By the covering number argument in Lemma A.7, we have

$$N(\epsilon, \mathcal{P}, l_2) \leq \left( C_1 / \epsilon \right)^{R^{D+1} + R \sum_{i=1}^{D} p_i + RK},$$

where $C_1 = 3D + 4$ is a constant. Suppose $a \in \mathcal{P}$ and $x \in \mathcal{N}(0, I_{s \times s})$, then by the Dudley’s integral entropy bound (see, e.g., Theorem 3.1 of Koltchinskii, 2011), we obtain

$$\mathbb{E}_x \sup_{a \in \mathcal{P}} (a^T x) \leq C_3 \int_0^1 \left\{ \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \log(C_1 / x) \right\}^{1/2} dx \leq C \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}.$$

Thus we complete the proof. \qed

**Lemma A.4.** Let $A \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K}$ and suppose $\mathcal{P}$ is defined in (A.50). Under Assumptions 2 and 5, we have

$$\sup_{\text{vec}(A) \in \mathcal{P}} \left| \frac{1}{n} \mathbb{E}\left\{ \frac{1}{n} \sum_{i=1}^{n} (A, \Phi(X_i))^2 \right\} - 1 \right| \leq C_1 \tilde{h}_n h_n^{-1} w(\mathcal{P}) \sqrt{n} \tag{A.52}$$

with probability at least $1 - \exp\{-C_2 w^2(\mathcal{P})\}$, where $w(\mathcal{P})$ is the Gaussian width, $(\Phi(X))_{j,k} = b_k(X_j)$ for $j \in J$, $k = 1, \ldots, K$, $J$ is defined in (A.1) and $h_n$ is defined in (A.3). Furthermore, suppose $n > C h_n^2 h_n^{-2} w^2(\mathcal{P})$ for some $C > 0$, then with the same probability, we have
\[ C_3 h_n \leq \inf_{\text{vec}(A) \in P} \frac{1}{n} \left| \sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \right|^2 \leq \sup_{\text{vec}(A) \in P} \frac{1}{n} \left| \sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \right|^2 \leq C_4 h_n. \quad (A.53) \]

Note that the above \( w(P) \) can be replaced by a constant \( t \), provided \( t \geq w(P) \) and \( P \) can be replaced by a subset of it.

**Proof.** Based on Lemma A.2, the following proof is similar to Theorem 12 of Banerjee et al. (2015). We consider the following class of functions

\[ F = \left\{ f_A : f_A \{ \Phi(X) \} = \frac{1}{\sqrt{E\{|\langle A, \Phi(X) \rangle|^2\}}} \langle A, \Phi(X) \rangle, \text{vec}(A) \in P \right\}. \]

It is trivial to see that \( F \subset S_{L_2} := \{ f : E[f^2(\Phi(X))] = 1 \} \). By definition,

\[ \sup_{f_A \in F} \| f_A \|_{\psi_2} = \sup_{\text{vec}(A) \in P} \left\| \frac{1}{\sqrt{E\{|\langle A, \Phi(X) \rangle|^2\}}} \langle A, \Phi(X) \rangle \right\|_{\psi_2}, \]

and by Lemma A.2, for every \( \text{vec}(A) \in P \),

\[ \left\| \frac{1}{\sqrt{E\{|\langle A, \Phi(X) \rangle|^2\}}} \langle A, \Phi(X) \rangle \right\|_{\psi_2} \leq \kappa_n, \]

where \( \kappa_n = C_5 h_n^{1/2} h_n^{-1/2} \). Then we obtain

\[ \sup_{f_A \in F} \| f_A \|_{\psi_2} \leq \kappa_n. \]

Thus for the \( \gamma_2 \) functionals, we have

\[ \gamma_2 (F \cap S_{L_2}, \| \cdot \|_{\psi_2}) \leq \kappa_n \gamma_2 (F \cap S_{L_2}, \| \cdot \|_{L_2}) \leq C_6 \kappa_n w(P), \]

where the last inequality follows from Theorem 2.1.1 of Talagrand (2005). By Theorem 10 of Banerjee et al. (2015), we can choose

\[ \theta = C_7 C_6 \kappa_n^2 \frac{w(P)}{\sqrt{n}} \geq C_7 \kappa_n \frac{\gamma_2 (F \cap S_{L_2}, \| \cdot \|_{\psi_2})}{\sqrt{n}}. \]

As a result, with probability at least \( 1 - \exp\{-C_8 \theta^2 n/\kappa_1^4\} \), we have (A.52) holds with \( C_1 = C_7 C_6 C_5^2 \) and \( C_2 = C_8 C_7^2 C_6^2 \). Suppose \( \sqrt{n} > C h_n h_n^{-1} w(P) \) for some \( C > 0 \), then by Lemma A.2, with probability at least \( 1 - \exp\{-C_2 w^2(P)\} \), we have

\[ C_3 h_n \leq \inf_{\text{vec}(A) \in P} \frac{1}{n} \left| \sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \right|^2 \leq \sup_{\text{vec}(A) \in P} \frac{1}{n} \left| \sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \right|^2 \leq C_4 h_n, \]

which completes the proof of (A.53). \( \square \)
Lemma A.5. Suppose \( A \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K} \), \( \text{rank}(A) \leq R \) and \( \sum_{k=1}^{K} A_{j,k} u_k = 0 \) for \( j \in J \setminus \{(1, \ldots, 1)\} \), where \( u_k = \int_0^1 b_k(x) \, dx \) and \( J \) is defined in (A.1). Under Assumptions 2, 3, and 5, if \( n > C \tilde{h}_n^2 h_n^{-2} (R^{D+1} + R \sum_{i=1}^{D} p_i + RK) \) for some constant \( C > 0 \), where \( h_n \) is defined in (A.3), we then have

\[
\sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \epsilon_i \leq C_1 \| A \|_{HS} \left\{ nh_n \left( R^{D+1} + \sum_{i=1}^{D} R p_i + RK \right) \right\}^{1/2},
\]

(A.54)

with probability at least

\[
1 - C_2 \exp \left\{ - C_3 \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.
\]

Proof. We use the notation \( Z = (z_1, \ldots, z_n)^T \) introduced in (A.12), then the left hand side of (A.54) can be rewritten as

\[
\sum_{i=1}^{n} \langle A, \Phi(X_i) \rangle \epsilon_i = (Za)^T \epsilon.
\]

Consider

\[
\Gamma_1 = \left\{ \frac{Za}{\sqrt{\lambda_{\max}(Z^T Z)}} : a \in \mathcal{P} \right\},
\]

where \( \lambda_{\max}(Z^T Z) = \sup_{a \in \mathcal{P}} \| Za \|_2 \) and \( \mathcal{P} \) is defined as (A.50). By the covering number argument in Lemma A.7,

\[
N(\epsilon, \mathcal{P}, l_2) \leq (C_4/\epsilon)^{R^{D+1} + R \sum_{i=1}^{D} p_i + RK},
\]

where \( C_4 = 3D + 4 \) is a constant. Following from the definition of \( \Gamma_1 \), we have

\[
N(\epsilon, \Gamma_1, l_2) \leq N(\epsilon, \mathcal{P}, l_2) \leq (C_4/\epsilon)^{R^{D+1} + R \sum_{i=1}^{D} p_i + RK}.
\]

By Assumption 3, for \( \eta \in \Gamma_1 \), \( \mathbb{E}\{\exp(t \eta^T \epsilon)\} \leq \exp(Ct^2 \| \eta \|^2) \leq \exp(Ct^2) \). Using the Dudley’s integral entropy bound, we have

\[
\mathbb{E} \sup_{\eta \in \Gamma_1} (\eta^T \epsilon) \leq C \int_0^2 \left\{ \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \log \left( C_4/\epsilon \right) \right\}^{1/2} \, d\epsilon
\]

\[
\leq C_5 \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2}.
\]

As a direct result (e.g., Theorem 8.1.6 of Vershynin, 2018), we have

\[
\sup_{\eta \in \Gamma_1} (\eta^T \epsilon) \leq C \left[ \int_0^2 \left\{ \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \log \left( C_4/\epsilon \right) \right\}^{1/2} \, d\epsilon + 2t \right]
\]

\[
\leq C_6 \left\{ \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2} + t \right\},
\]

44
with probability at least $1 - 2 \exp(-t^2)$, which implies

$$
(Za)^T \epsilon \leq C_7 \sqrt{\lambda_{\text{Rmax}}(Z^T Z)} \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right)^{1/2},
$$

(A.55)

with probability at least

$$
1 - 2 \exp\left\{ - \left( R^{D+1} + R \sum_{i=1}^{D} p_i + RK \right) \right\}.
$$

Plugging (A.51) and (A.53) into (A.55), we will complete the proof of (A.54).

\[ \square \]

**Lemma A.6.** Suppose $\int_0^1 f_r(u) du = 0$, $r = 1, \ldots, R$. If Assumptions 4 and 5 hold, then there exist $\alpha_{0r,k}$, $k = 1, \ldots, K$, such that

$$
\| f_r - \sum_{k=1}^{K} \alpha_{0r,k} b_k \|_\infty = \mathcal{O}(K^{-\tau}),
$$

where $\sum_{k=1}^{K} \alpha_{0r,k} u_k = 0$ and $u_k = \int_0^1 b_k(x) dx$.

**Proof.** It follows from Assumptions 4, 5, and Lemma 5 of Stone (1985) that that for each $r$, there exists a spline function $f_{1r}$ which can be represented by $\{b_k(x)\}_{k=1}^{K}$, such that

$$
\| f_r - f_{1r} \|_\infty = \mathcal{O}(K^{-\tau}).
$$

Let $f_{2r} = f_{1r} - \int_0^1 f_{1r}(u) du$, then we have

$$
\| f_r - f_{2r} \|_\infty \leq \| f_r - f_{1r} \|_\infty + \left| \int_0^1 f_{1r}(u) du \right|.
$$

Since

$$
\left| \int_0^1 f_{1r}(u) du \right| = \left| \int_0^1 \{ f_{1r}(u) - f(u) \} du \right| + \int_0^1 f(u) du
\leq \| f_r - f_{1r} \|_\infty
= \mathcal{O}(K^{-\tau}),
$$

it is straightforward to get

$$
\| f_r - f_{2r} \|_\infty = \mathcal{O}(K^{-\tau}).
$$

The proof is completed by noting that $f_{2r}$ is a spline function with mean zero.

\[ \square \]

**Lemma A.7.** Let $A \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K}$. To simplify the notations, denote $p_{D+1} = K$. Let

$$
\Gamma_2 = \{ a : \| a \|_2 \leq 1, a = \text{vec}(A), \text{rank}(A) \leq R \}.
$$

Then the covering number of $\Gamma_2$ satisfies

$$
N(\epsilon, \Gamma_2, l_2) \leq \left( \frac{3D + 4}{\epsilon} \right)^{R^{D+1} + R \sum_{d=1}^{D+1} p_d}.
$$

(A.56)
Proof. Since the CP decomposition is a special case of the Tucker decomposition (Kolda and Bader, 2009), \( \mathbf{A} \) can be represented as

\[
\mathbf{A} = \mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{B}_{D+1},
\]

(A.57)

where \( \mathbf{I} \in \mathbb{R}^{R \times R \times \cdots \times R} \) is a diagonal tensor of which all the diagonal entries are 1, \( \mathbf{B}_d \in \mathbb{R}^{p_d \times R} \), and \( \times_d \) denotes the \( d \)-mode (matrix) product of a tensor with a matrix (Kolda and Bader, 2009). Let \( r_d = \text{rank}(\mathbf{B}_d) \). Through the QR decomposition, we get \( \mathbf{B}_d = \mathbf{Q}_d \mathbf{R}_d \), where \( \mathbf{Q}_d \mathbf{I}_d = \mathbf{I}_r \) with \( \mathbf{I}_d \in \mathbb{R}^{r_d \times r_d} \) the identity matrix. Applying the argument to (A.57), we have

\[
\mathbf{A} = (\mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D) \times_{D+1} (\mathbf{Q}_D \mathbf{R}_{D+1})
\]

(A.58)

\[= \mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{Q}_D \mathbf{R}_{D+1} \]

In other words, the CP decomposition will lead a higher-order singular value decomposition (HOSVD) (see, e.g., De Lathauwer et al., 2000). By Lemma 2 of Rauhut et al. (2017), we obtain

\[
N(\epsilon, \Gamma_2, l_2) \leq \left( \frac{3D + 4}{\epsilon} \right)^{\sum_{d=1}^{D+1} r_d + 1}\frac{1}{\epsilon}.
\]

Therefore (A.56) is shown by noting that \( r_d \leq R \) for \( d = 1, \ldots, D + 1 \).

Lemma A.8. Represent the CP low-rank tensor \( \mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_{D+1}} \) as in (A.57), where \( R \) is its CP-rank and \( \mathbf{B}_d \in \mathbb{R}^{p_d \times R} \), \( d = 1, \ldots, D + 1 \). Let \( \delta \geq 0 \) and

\[
\Gamma_\delta = \{ \mathbf{A} : \mathbf{A} = \mathbf{I} \times_1 \mathbf{B}_1 \times_2 \cdots \times_D \mathbf{B}_D \times_{D+1} \mathbf{B}_{D+1}, \| \mathbf{B}_d \|_2 \leq \delta, d = 1, \ldots, D + 1, \| \mathbf{A} \| \leq 1 \}.
\]

Then, the covering number of \( \Gamma_\delta \) satisfies

\[
N(\epsilon, \Gamma_\delta, l_2) \leq \left( \frac{C_1}{\epsilon} \right) R^2 \log \epsilon \leq \epsilon \sum_{d=1}^{D+1} p_d
\]

(A.59)

Proof. Using (A.58) in the proof of Lemma A.7, \( \mathbf{A} \) can be written as

\[
\mathbf{A} = (\mathbf{I} \times_1 \mathbf{R}_1 \times_2 \cdots \times_D \mathbf{R}_{D+1}) \times_1 \mathbf{Q}_1 \times_2 \cdots \times_{D+1} \mathbf{Q}_{D+1},
\]

where \( \mathbf{R}_d \in \mathbb{R}^{R \times R} \), \( \mathbf{Q}_d \in \mathbb{R}^{p_d \times R} \) with \( \mathbf{Q}_d^T \mathbf{Q}_d = \mathbf{I}_R \), and \( \mathbf{Q}_d \mathbf{R}_d = \mathbf{B}_d \). The above HOSVD representation is not unique since respectively replacing \( \mathbf{R}_d \) and \( \mathbf{Q}_d \) by \( \mathbf{U}_d^T \mathbf{R}_d \) and \( \mathbf{Q}_d \mathbf{U}_d \) shares the equivalent meaning of HOSVD, for any \( \mathbb{R}^{R \times R} \) orthonormal matrix \( \mathbf{U}_d \). In order to use the argument of Rauhut et al. (2017), denote

\[
\mathbf{S} = \mathbf{I} \times_1 \mathbf{R}_1 \times_2 \cdots \times_D \mathbf{R}_d \times_{D+1} \cdots \times_{D+1} \mathbf{R}_{D+1}.
\]

(A.60)
We first show that there exists a HOSVD representation such that the matrix SVD of \( S_{(d)} \) satisfies the special form of

\[
S_{(d)} = \Sigma_d V_d^T, \quad d = 1, \ldots, D + 1, \quad (A.61)
\]

where \( S_{(d)} \) is the mode-\( d \) matricization of \( S \) defined in (A.60). In the following, we construct a HOSVD representation from its equivalent class. Indeed, the general matrix SVD of \( S_{(d)}; d = 1, \ldots, D + 1 \), can be written as

\[
S_{(d)} = U_d \Sigma_d V_d^T,
\]

where \( U_d U_d^T = I_R, \ V_d^T V_d = I_R \), and \( \Sigma_d \) is a diagonal matrix. Let

\[
\tilde{S} = S \times_1 U_1^T \times_2 \cdots \times_{D+1} U_{D+1}^T = I \times_1 U_1^T R_1 \times_2 \cdots \times_{D+1} U_{D+1}^T R_{D+1}.
\]

The HOSVD of \( A \) can then be written as

\[
A = \tilde{S} \times_1 Q_1 \times_2 \cdots \times Q_{D+1} U_{D+1},
\]

where \( (Q_d U_d)^T Q_d U_d = I_R \) and \( \|U_d^T R_d\| = \|R_d\| = \|Q_d R_d\| = \|B_d\| \leq \delta \). It further shows that matrix SVD of \( \tilde{S}_{(d)} \) has the form of

\[
\tilde{S}_{(d)} = \Sigma_d V_d^T (U_{D+1}^T \otimes \cdots \otimes U_{d+1}^T \otimes U_{d-1}^T \otimes \cdots \otimes U_1^T)^T,
\]

where \( (U_{D+1}^T \otimes \cdots \otimes U_{d+1}^T \otimes U_{d-1}^T \otimes \cdots \otimes U_1^T) V_d \) is orthonormal. The above discussion concludes that any \( A \in \Gamma_\delta \) can be represented as

\[
A = S \times_1 Q_1 \times_2 \cdots \times Q_{D+1},
\]

where \( S \) is defined as in (A.60) and satisfies (A.61).

The definition of \( S \) in (A.60) implies that \( \text{rank}(S) \leq R \). We next consider the \( \epsilon \)-net of

\[
S = \{S, \|S\| \leq 1, \text{rank}(S) \leq R, \|R_d\|^2 \leq \delta \}.
\]

Suppose \( S, \tilde{S} \in S \), and by definition they can be written as

\[
S = I_1 \times_1 R_1 \times_2 \cdots \times_d R_d \times_{d+1} \cdots \times_{D+1} R_{D+1}
\]

and

\[
\tilde{S} = I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d \bar{R}_d \times_{d+1} \cdots \times_{D+1} \bar{R}_{D+1}
\]

for some \( R_d, \bar{R}_d \in \mathbb{R}^{R \times R} \), respectively. Note that

\[
\|S - \tilde{S}\| = \|I_1 \times_1 R_1 \times_2 \cdots \times_d R_d \times_{d+1} R_{D+1} - I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d \bar{R}_d \times_{d+1} R_{D+1}\|
\]

\[
= \|I_1 \times_1 R_1 \times_2 \cdots \times_d R_d \times_{d+1} R_{D+1} - I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d \bar{R}_d \times_{d+1} R_{D+1}\|
\]

\[
+ I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d \bar{R}_d \times_{d+1} R_{D+1} - I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d \bar{R}_d \times_{d+1} \bar{R}_{D+1}\|
\]

\[
+ I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d R_d \times_{d+1} \bar{R}_{D+1} - I_1 \times_1 \bar{R}_1 \times_2 \cdots \times_d R_d \times_{d+1} \bar{R}_{D+1}\|
\]

\[
\leq \sum_{d=1}^{D+1} \|I_1 \times_1 R_1 \times_2 \cdots (R_d - \bar{R}_d) \cdots \times_d R_d \times_{d+1} R_{D+1}\|.
\]

47
Using the Cauchy-Swarchz inequality, we obtain

\[
\| \mathbf{I} \times_1 \mathbf{R}_1 \times_2 \cdots (\mathbf{R}_d - \bar{\mathbf{R}}_d) \cdots \times_D \mathbf{R}_D \times_{D+1} \mathbf{R}_{D+1} \|_2^2
\]

\[
= \sum_{t,v=1}^R \sum_{i_1, \ldots, i_{D+1}} R_{1,i_1,t} R_{1,i_1,v} \cdots (R_{d,i_d,t} - \bar{R}_{d,i_d,t})(R_{d,i_d,v} - \bar{R}_{d,i_d,v}) \cdots \bar{R}_{D+1,i_{D+1},t} \bar{R}_{D+1,i_{D+1},v}
\]

\[
\leq \sum_{t,v=1}^R \left( \sum_{i_1} R_{1,i_1,t} \sum_{i_1} R_{1,i_1,v} \right)^{0.5} \cdots \left( \sum_{i_{D+1}} R_{1,i_{D+1},t} \sum_{i_{D+1}} R_{1,i_{D+1},v} \right)^{0.5}
\]

\[
\leq \delta^D \sum_{t,v} \left\{ \sum_{i_d} (R_{d,i_d,t} - \bar{R}_{d,i_d,t})^2 + \sum_{i_d} (R_{d,i_d,v} - \bar{R}_{d,i_d,v})^2 \right\}
\]

\[
\leq CR\delta^D \| \mathbf{R}_d - \bar{\mathbf{R}}_d \|_2^2,
\]

where \( R_{d,i_1,i_2} \) and \( \bar{R}_{d,i_1,i_2} \) are the \((l_1, l_2)\)-th element of \( \mathbf{R} \) and \( \bar{\mathbf{R}} \), respectively. Since the \( \epsilon/(C\delta^D) \)-net of a matrix space \( \mathcal{R} = \{ \mathbf{R} : \mathbf{R} \in \mathbb{R}^{R \times R}, \| \mathbf{R} \| \leq 1 \} \) is \( (C\delta^D/\epsilon)^R \), using the similar arguments as in the proof of Lemma 2 in Rauhut et al. (2017), the \( \epsilon \) net of \( \mathcal{S} \) is bounded by

\[
\left( \frac{C_3}{\epsilon} \right) C_4 R^2 \log R \delta.
\]

(A.62)

Following the arguments in Rauhut et al. (2017) with (A.61) and using (A.62), we get the result of (A.59).

Lemma A.9. Suppose \((\hat{\mathbf{A}}_{\text{PLS}}, \hat{\nu}_{\text{PLS}})\) is a solution to (10) and

\[
\hat{\mathbf{A}}_{\text{PLS}} = \sum_{r=1}^R \hat{\alpha}_r \circ \hat{\beta}_r \circ \cdots \circ \hat{\beta}_r \circ \hat{\alpha}_r,
\]

where \( \hat{\beta}_r \) and \( \hat{\alpha}_r \) are the corresponding parameterization to make the penalty of (10) the smallest. If Assumptions 1 and 3 hold, then

\[
\sum_{d=1}^D \sum_{r=1}^R \| \hat{\beta}_{r,d} \|^2 \leq \delta_B,
\]

(A.63)

and

\[
\hat{\nu}_{\text{PLS}} \leq \frac{C_1}{s} \sum_{r=1}^{R_0} \| \text{vec}(\mathbf{B}_{0r}) \|_1 + |\nu_0| + C_2 + C_3 \sqrt{K} \delta_B D/\sqrt{s},
\]

(A.64)

with probability at least \( 1 - C_4 \exp(-C_5 n) \), where

\[
\delta_B = \frac{1}{\varpi(n)} \left\{ \left( \sum_{r=1}^{R_0} \| \text{vec}(\mathbf{B}_{0r}) \|_1 \right)^2 \frac{C_6}{K^2} + C_7 n + G_0 \right\},
\]

and \( G_0 \) is defined in (A.9).
Proof. It follows from Assumption 3 that \( \epsilon_i \) is sub-Gaussian. By Lemma 2.7.6 and Exercise 2.7.10 of Vershynin (2018), \( \epsilon^2_i \) and \( \epsilon^2_i - \mathbb{E}\epsilon^2_i \) are sub-exponential. Using Corollary 2.8.3 of Vershynin (2018),

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \{ \epsilon^2_i - \mathbb{E}(\epsilon^2_i) \} \right| \leq t,
\]

with probability at least

\[
1 - 2 \exp\left\{ - C_1 \min\left( \frac{t^2}{C_2^2}; \frac{t}{C_2} \right) n \right\}.
\]

Taking \( t = C_1 \) yields

\[
\sum_{i=1}^{n} \{ \epsilon^2_i - \mathbb{E}(\epsilon^2_i) \} \leq C_1 n,
\]

with probability at least \( 1 - 2 \exp(-C_2 n) \). By Proposition 2.7.1 of Vershynin (2018), we obtain \( \mathbb{E}(\epsilon^2_i) \leq C \), which yields

\[
\sum_{i=1}^{n} \epsilon^2_i \leq C_1 n, \tag{A.65}
\]

with probability at least \( 1 - 2 \exp(-C_2 n) \). Using (A.19), (A.10) and (A.65), we have

\[
\hat{G} \leq \sum_{i=1}^{n} \left( y_i - \nu_0 - \frac{1}{s} \langle A_0, \Phi(X_i) \rangle \right)^2 + G_0
\]

\[
\leq \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(B_0r) \|_1}{s} \right\}^2 \frac{C_1}{K^{2\tau}} + 2 \sum_{i=1}^{n} \epsilon^2_i + G_0 \tag{A.66}
\]

\[
\leq \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(B_0r) \|_1}{s} \right\}^2 \frac{C_1}{K^{2\tau}} + C_2 n + G_0,
\]

with probability at least \( 1 - 2 \exp(-C n) \), where \( \hat{G} \) is defined in (A.5). Let

\[
\delta_p = \left\{ \sum_{r=1}^{R_0} \frac{\| \text{vec}(B_0r) \|_1}{s} \right\}^2 \frac{C_1}{K^{2\tau}} + C_2 n + G_0 \right\}^{S_G}.
\]

By Assumption 1, (A.66) and the definition of \( \delta_p \), we have \( \hat{G}^{S_G} \leq \delta_p \), and

\[
\sum_{d=1}^{D} \sum_{r=1}^{R} \| \hat{\beta}_{r,d} \|_2^2 \leq \frac{\delta_p}{\varpi(n)}, \tag{A.67}
\]

with probability at least \( 1 - 2 \exp(-C n) \), which completes the proof of (A.63). By (A.67), we obtain

\[
\| \hat{\beta}_{r,d} \|_2^2 \leq \frac{\delta_p}{\varpi(n)}.
\]

49
Thus,
\[
\|\hat{A}\| = \left\| \sum_{r=1}^{R} \hat{\beta}_{r,1} \circ \cdots \circ \hat{\beta}_{r,D} \circ \hat{\alpha}_r \right\| \\
\leq \sum_{r=1}^{R} \|\hat{\beta}_{r,1} \circ \cdots \circ \hat{\beta}_{r,D} \circ \hat{\alpha}_r\| \\
= \sum_{r=1}^{R} \|\hat{\beta}_{r,1}\|_2 \cdots \|\hat{\beta}_{r,D}\|_2 \|\hat{\alpha}_r\|_2 \\
\leq R \left\{ \frac{\delta_p}{\varpi(n)} \right\}^{D/2}.
\]

Note that \(\|\tilde{\Phi}(X_i)\|^2 \leq KsC\) and
\[
|y_i - \epsilon_i| = |\nu_0 + \frac{1}{s} \sum_{r=1}^{R_0} (B_{0r}, F_r(X_i))| \\
\leq \frac{1}{s} \sum_{r=1}^{R_0} |(B_{0r}, F_r(X_i))| + |\nu_0| \\
\leq \frac{C_1}{s} \sum_{r=1}^{R_0} \|B_{0r}\|_1 + |\nu_0|.
\]

We then have
\[
\hat{\nu}_{\text{PLS}} = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \frac{1}{s} \langle \hat{A}, \tilde{\Phi}(X_i) \rangle \right\} \\
\leq \left| \frac{1}{n} \sum_{i=1}^{n} (y_i - \epsilon_i) \right| + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i + \frac{1}{sn} \sum_{i=1}^{n} \|\hat{A}\| \|\tilde{\Phi}(X_i)\| \\
\leq \left| \frac{1}{n} \sum_{i=1}^{n} (y_i - \epsilon_i) \right| + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i + \frac{1}{sn} \sum_{i=1}^{n} \|\hat{A}\| \|\tilde{\Phi}(X_i)\| \\
\leq \frac{C_1}{s} \sum_{r=1}^{R_0} \|B_{0r}\|_1 + |\nu_0| + C_2 + C_3 \sqrt{K} R \left\{ \frac{\delta_p}{\varpi(n)} \right\}^{D/2} / \sqrt{s},
\]
with probability at least \(1 - C_4 \exp(-C_5 n)\), which completes the proof of (A.64). \(\square\)

**Lemma A.10.** If \(G(\theta)\) is defined as (12) with (14), \(\lambda_1 > 0\), and \(0 \leq \lambda_2 \leq 1\), then \(G(\theta)\) satisfies Assumption 1 with \(S_G = 1\) or \(S_G = 2\) and
\[
\varpi(n) = \frac{1}{2} (1 - \lambda_2)^2 \lambda_1 \quad \text{or} \quad \lambda_1^2 \lambda_2^2,
\]
where \(\theta\) is defined in (11).

**Proof.** By definition
\[
G(\theta) = \lambda_1 \sum_{d=1}^{D} \sum_{r=1}^{R} \left\{ \frac{1}{2} (1 - \lambda_2) \|\beta_{r,d}\|^2 + \lambda_2 \|\beta_{r,d}\|_1 \right\}.
\]

50
If $0 \leq \lambda_2 < 1$, then
\[
G(\theta) \geq \frac{1}{2}(1 - \lambda_2)\lambda_1 \sum_{d=1}^{D} \sum_{r=1}^{R} \|\beta_{r,d}\|^2.
\]
If $\lambda_2 = 1$, then
\[
G(\theta) \geq \lambda_1 \lambda_2 \sum_{d=1}^{D} \sum_{r=1}^{R} \|\beta_{r,d}\|_1 \geq \lambda_1 \lambda_2 \left( \sum_{d=1}^{D} \sum_{r=1}^{R} \|\beta_{r,d}\|^2 \right)^{1/2}.
\]
Therefore, we finish the proof.

Before presenting Lemma A.11, we introduce some notations. Recall that $\nu_0$ and
\[
A_0 = \sum_{r=1}^{R_0} \beta_{0r,1} \circ \cdots \circ \beta_{0r,D} \circ \alpha_{0r},
\]
where $\alpha_{0r}^T(u_1, \ldots, u_K)^T = 0$ and $u_k = \int_0^1 b_k(x)dx$, correspond to the best approximation model (5). By the equivalence of basis (Ruppert et al., 2003) and Lemma B.1 in Subsection B.1, there exists a matrix $Q$ such that
\[
Qb(t) = \tilde{b}(t) \quad \text{for all} \quad t.
\] (A.68)

Let
\[
\tilde{\alpha}_{0r}^T = (\tilde{\alpha}_{0r,1}, \ldots, \tilde{\alpha}_{0r,K-1}, \tilde{\alpha}_{0r,K}) = \alpha_{0r}^T Q^{-1},
\]
and
\[
\delta_{0r} = \sum_{k=1}^{K} \tilde{\alpha}_{0r,k} \tilde{u}_k, \quad \tilde{u}_k = \int_0^1 \tilde{b}(x)dx.
\]

Suppose
\[
\tilde{\alpha}_{0r} = (\tilde{\alpha}_{0r,1}, \ldots, \tilde{\alpha}_{0r,K-1})^T \in \mathbb{R}^{K-1},
\]
and its entry is defined by
\[
\tilde{\alpha}_{0r,k} = \tilde{\alpha}_{0r,k+1} \quad \text{for} \quad k = 1, \ldots, K - 1.
\]

We further let
\[
\tilde{\nu}_0 = \nu_0 + \frac{1}{s} \sum_{r=1}^{R_0} \left( \beta_{0r,1} \circ \cdots \circ \beta_{0r,D} \circ \delta_{0r} \right),
\]
and
\[
\tilde{A}_0 = \sum_{r=1}^{R_0} \|\tilde{\alpha}_{0r}\|^{1/D} \beta_{0r,1} \circ \cdots \circ \tilde{\alpha}_{0r} \|^{1/D} \beta_{0r,D} \circ \|\tilde{\alpha}_{0r}\|,
\]
where $J \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is the tensor with all the entries to be 1. It can be shown that the regression function $m(\cdot)$ remains invariant using the above basis transformation, i.e.,
\[
\nu_0 + \frac{1}{s} \langle A_0, \Phi(X_i) \rangle = \tilde{\nu}_0 + \frac{1}{s} \langle \tilde{A}_0, \tilde{\Phi}(X_i) \rangle.
\]
Later we will use two conditions, i.e.,

$$|\tilde{\nu}_0| \leq \delta_v,$$  \hspace{1cm} (A.69)

and

$$\sum_{r=1}^{R} \|\hat{\alpha}_{0r}\|^{2/D} \|\beta_{0r,d}\|^{2} \leq \delta_B,$$  \hspace{1cm} (A.70)

**Lemma A.11.** Consider the constrained optimization

$$\arg \min_{\tilde{\nu}, \hat{A}} \sum_{i=1}^{n} \left( y_i - \tilde{\nu} - \frac{1}{s} \langle \hat{A}, \tilde{\Phi}(X_i) \rangle \right)^2$$

s.t. \hspace{1cm} \hat{A} = \sum_{r=1}^{R} \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \hat{\alpha}_r,

\hspace{1cm} \sum_{r=1}^{R} \|\beta_{r,d}\|_2 \leq \delta_B, \text{ for } d = 1, \ldots, D,

\hspace{1cm} |\tilde{\nu}| \leq \delta_v,

\hspace{1cm} \|\hat{\alpha}_{r}\|_2^2 = 1, \text{ for } r = 1, \ldots, R. \hspace{1cm} (A.71)

Suppose \(\delta_v \) and \(\delta_B \) satisfy (A.69) and (A.70), respectively. Let \(\hat{m}_{\text{con}}(X)\) be the estimated function constructed from the aforementioned optimization (A.71). If \(n > C_1 \tilde{h}_0^2 h_n^{-2} (R^2 \log \delta_{\text{con}} + R \sum_{d=1}^{D} p_d)\), we have

$$\|\hat{m}_{\text{con}} - m_0\|^2 \leq C_2 \frac{R^2 \log \delta_{\text{con}} + R \sum_{d=1}^{D} p_d + RK}{n} + C_3 \left\{ \frac{\sum_{r=1}^{R_0} \|\text{vec}(B_{0r})\|_1}{s} \right\}^2 \frac{1}{K^{2r}},$$

with probability at least

$$1 - C_4 \exp \left\{ - C_5 \left( R^2 \log \delta_{\text{con}} + R \sum_{d=1}^{D} p_d + RK \right) \right\},$$

where

$$\delta_{\text{con}} = n \max \left\{ C_6 (\delta_v s + \sqrt{s} R^2 \delta_B^{D/2})^{2/D}, C_7 RK^2, \sum_{r=1}^{R_0} \|\beta_{0r,d}\|_2^2 + (s \nu_0)^2/D, \sum_{r=1}^{R_0} \|\alpha_{0r}\|_2^2 + 1 \right\}.$$  \hspace{1cm} (A.76)

**Proof.** The proof includes three parts. First, we will show the an inequality to upper bound the in-sample error (see (A.76)). Suppose \( (\hat{\nu}_{\text{con}}, \hat{A}_{\text{con}}) \) is a solution of (A.71), where

$$\hat{A}_{\text{con}} = \sum_{r=1}^{R} \hat{\beta}_{r,1} \circ \hat{\beta}_{r,2} \circ \cdots \circ \hat{\beta}_{r,D} \circ \hat{\alpha}_r \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K-1}.$$  \hspace{1cm} (A.77)

Using (A.69) and (A.70), we have

$$\sum_{i=1}^{n} \left( y_i - \hat{\nu}_{\text{con}} - \frac{1}{s} \langle \hat{A}_{\text{con}}, \tilde{\Phi}(X_i) \rangle \right)^2 \leq \sum_{i=1}^{n} \left( y_i - \nu_0 - \frac{1}{s} \langle A_0, \Phi(X_i) \rangle \right)^2.$$  \hspace{1cm} (A.72)
To present the proof clearly, we let
\[ \tilde{\beta}_{r,1} = \tilde{\beta}_{r,2} \circ \cdots \circ \tilde{\beta}_{r,D}, \]
\[ \alpha_{\text{tem},r} = (\alpha_{\text{tem},r,1}, \ldots, \alpha_{\text{tem},r,K})^T, \]
\[ \alpha_{\text{tem},r,k} = \hat{\alpha}_{r,k-1}, \text{ for } k = 2, 3, \ldots, K, \]
\[ \alpha_{\text{tem},r,1} = -\delta_r, \]
where \( \delta_r = \sum_{k=1}^{K-1} \hat{\alpha}_{r,k} \tilde{u}_{k+1} \) and \( \tilde{u}_k = \int_0^1 \tilde{b}(x) dx \). We further denote
\[ A_{\text{tem}} = \sum_{r=1}^R \tilde{\beta}_{r,1} \circ \alpha_{\text{tem},r}, \]
and
\[ \nu_{\text{tem}} = \nu_{\text{con}} + \frac{1}{s} \sum_{r=1}^R \langle \tilde{\beta}_{r,1}, \frac{\delta_r}{\tilde{u}_1} \rangle. \]

By their definitions, we have \( \sum_{k=1}^K \alpha_{\text{tem},r,k} \tilde{u}_k = 0 \) and
\[ \nu_{\text{tem}} + \frac{1}{s} \langle A_{\text{tem}}, \tilde{\Phi}(X) \rangle = \nu_{\text{con}} + \frac{1}{s} \langle \hat{A}_{\text{PLS}}, \tilde{\Phi}(X) \rangle. \]

Let \( \tilde{\nu}_{\text{con}} = \nu_{\text{tem}} \), and \( \hat{A}_{\text{con}} = \sum_{r=1}^R \tilde{\beta}_{r,1} \circ \alpha_r \), where
\[ \alpha_r^T = \alpha_{\text{tem},r}^T Q, \quad \text{(A.73)} \]
and \( Q \) is defined in (A.68). It further shows that,
\[ \hat{A}_{\text{con}} = \sum_{r=1}^R \tilde{\beta}_{r,1} \circ \tilde{\beta}_{r,2} \circ \cdots \circ \tilde{\beta}_{r,D} \circ \alpha_r \in \mathbb{R}^{p_1 \times \cdots \times p_D \times K}, \]
\[ \nu_{\text{con}} + \frac{1}{s} \langle \hat{A}_{\text{con}}, \tilde{\Phi}(X) \rangle = \hat{\nu}_{\text{con}} + \frac{1}{s} \langle \hat{A}_{\text{PLS}}, \tilde{\Phi}(X) \rangle, \quad \text{(A.74)} \]
and the elements of \( \alpha_r \), denoted as \( \{\hat{\alpha}_{r,1}, \ldots, \hat{\alpha}_{r,K}\} \), satisfy \( \sum_{k=1}^K \hat{\alpha}_{r,k} u_k = 0 \), where \( u_k = \int_0^1 b_k(x) dx \). Using (A.74) and (A.72) yields
\[ \sum_{i=1}^n \left( y_i - \hat{\nu}_{\text{con}} - \frac{1}{s} \langle \hat{A}_{\text{con}}, \tilde{\Phi}(X_i) \rangle \right)^2 \leq \sum_{i=1}^n \left( y_i - \nu_0 - \frac{1}{s} \langle A_0, \tilde{\Phi}(X_i) \rangle \right)^2. \quad \text{(A.75)} \]

Similar to the proof in Theorem 1, we also write the “design” matrix as \( Z = (z_1, \ldots, z_n)^T \in \mathbb{R}^n \times sK \), \( z_i = \text{vec}\{\Phi(X_i)\} \), \( i = 1, \ldots, n \) and use the absorbing mapping to define \( \hat{A}_{\text{con}} := \Omega(\hat{A}_{\text{con}}, \tilde{\nu}_{\text{con}}), \quad A_0 := \Omega(A_0, \nu_0) \), where \( \Omega \) is defined in (A.2). For notational simplicity, let
\[ A_{\text{con}}^\sharp = \hat{A}_{\text{con}} - A_0^\sharp \]
and $\tilde{a}_{\text{con}}^{\circ} = \text{vec}(A_{\text{con}}^{\circ})$. Multiplying the non-negative function $1_{\|a_{\text{con}}^{\circ}\| \geq \gamma}$ on both sides of (A.75), it can be shown for all $\gamma \in \mathbb{R}$,

$$\frac{1}{s^2} \left\| Z_{\text{con}} \right\|_{2,\gamma}^2 \leq 2 \left\langle \frac{1}{s} Z_{\text{con}} a_{\text{con}}^{\circ} \right\| a_{\text{con}}^{\circ} \|_{\gamma}, \epsilon \right\rangle + 2 \left\langle \frac{1}{s} Z_{\text{con}} a_{\text{con}}^{\circ} \right\| a_{\text{con}}^{\circ} \|_{\gamma}, v - \epsilon - \frac{1}{s} Z_{\text{con}} \right\rangle.$$

(A.76)

Second, we will show there exists a CP decompositions of $A_{\text{con}}^{\circ}$ with bounded CP components. Using de Boor-Fix functional (de Boor and Fix, 1973), we have

$$|Q_{k_1,k_2}| \leq C,$$

(A.77)

where $Q_{k_1,k_2}$ is the $(k_1,k_2)$-th entry of $Q$. (A.73) and (A.77) yield

$$\|\tilde{\alpha}_r\|_2 \leq C_1 K^2.$$  

(A.78)

After noting that

$$\|\tilde{B}_{\text{con},r}\|^2 = \|\tilde{B}_{r,1}\|^2 \cdots \|\tilde{B}_{r,D}\|^2 \leq \delta_B^D,$$

we then have

$$\nu_{\text{con}} \leq \nu + C_2 R \delta_B^{D/2} \sqrt{s}.$$  

(A.79)

By (A.78), (A.79), and the definition of $\tilde{A}_{\text{con}}^{\circ}$, there exists a CP decomposition of $\tilde{A}_{\text{con}}^{\circ}$,

$$\tilde{A}_{\text{con}}^{\circ} = \sum_{r=1}^{R+1} \tilde{B}_{r,1} \circ \tilde{B}_{r,2} \circ \cdots \circ \tilde{B}_{r,D} \circ \tilde{\alpha}_r,$$

satisfying

$$\sum_{r=1}^{R+1} \|\tilde{B}_{r,d}\|^2 \leq \delta_B + (s \nu_{\text{con}})^{2/D} \leq C_3 (\nu s + \sqrt{s} R \delta_B^{D/2})^{2/D}$$

and

$$\sum_{r=1}^{R+1} \|\tilde{\alpha}_r\|_2^2 \leq C_4 R K^2.$$

Analogously, there exists a CP decomposition of $A_0^{\circ}$,

$$A_0^{\circ} = \sum_{r=1}^{R_0+1} \beta_{0,r,1} \circ \cdots \circ \beta_{0,r,D} \circ \alpha_{0,r},$$

satisfying

$$\sum_{r=1}^{R_0+1} \|\beta_{0,r,d}\|^2 \leq \sum_{r=1}^{R_0} \|\beta_{0,r,d}\|^2 + (s \nu_{0})^{2/D}$$

and

$$\sum_{r=1}^{R_0+1} \|\alpha_{0,r}\|_2^2 \leq \sum_{r=1}^{R_0} \|\alpha_{0,r}\|_2^2 + 1.$$

Denote

$$\delta_{sK_0} = \max \left\{ C_3 (\nu s + \sqrt{s} R \delta_B^{D/2})^{2/D}, C_4 R K^2, \sum_{r=1}^{R_0} \|\beta_{0,r,d}\|^2 + (s \nu_{0})^{2/D}, \sum_{r=1}^{R_0} \|\alpha_{0,r}\|_2^2 + 1 \right\},$$

and

$$\mathcal{P}_{\delta_{sK_0}} = \{ A : A = I \times_1 B_1 \cdots \times_{D+1} B_{D+1}, \|B_d\|^2 \leq \delta_{sK_0}, B_d \in \mathbb{R}^{p_d \times 1}, d = 1, \ldots, D + 1 \},$$

54
where $R_1 = R_0 + R + 2$. We then have

$$A_{\text{con}}^\sharp \in \mathcal{P}_\delta := \{ A : A \in \mathcal{P}_1 \cap \mathcal{P}_{\delta_0K_0} \},$$

where $\mathcal{P}_1$ is defined as (A.14). In other words, there exists a CP decomposition of $A_{\text{con}}^\sharp$ with bounded components.

Third, we will show the final result. Let $\gamma = 1/n$ and

$$\mathcal{P}_{\delta\gamma} = \left\{ \frac{A}{\|A\|} : A \in \mathcal{P}_\delta \cap \{ \|A\| \geq \gamma \} \right\}.$$

For all $A \in \mathcal{P}_{\delta\gamma}$, there exists one CP decomposition of $A$, 

$$A = I \times_1 B_1 \times \cdots \times D+1 B_{D+1},$$

satisfying

$$\|B_d\| \leq \frac{\delta_0K_0}{\gamma}, d = 1, \ldots, D + 1.$$

Using Dudley’s integral entropy bound (Vershynin, 2018) and Lemma A.8, we obtain

$$w^2(\mathcal{P}_\gamma) \leq C_8 R^2 \log \delta_0K_0/\gamma + C_9 R \sum_{d=1}^{D+1} p_d.$$  \hspace{1cm} (A.80)

Our assumption on $n$ and (A.80) imply that

$$C_5 n h_n \|a_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma} \leq \|Za_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma} \leq C_6 n h_n \|a_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma}^2,$$  \hspace{1cm} (A.81)

with probability at least

$$1 - 2 \exp \left\{ - C \left( R^2 \log \delta_0K_0/\gamma + R \sum_{d=1}^{D+1} p_d \right) \right\},$$

due to the similar arguments used in the proof of Lemma A.4. Using Dudley’s integral entropy bound again and Lemma A.8, we further have

$$\langle Za_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma}, \epsilon \rangle \leq C_2 \|a_{\text{con}}^\sharp 1\|_2 (n h_n)^{1/2} \left( R^2 \log R \delta_0K_0/\gamma + R \sum_{d=1}^{D+1} p_d \right)^{1/2}$$

with probability at least

$$1 - C_3 \exp \left\{ - C_4 \left( R^2 \log \delta_0K_0/\gamma + R \sum_{d=1}^{D+1} p_d \right) \right\},$$

as shown in the proof of Lemma A.5. Similar to (A.20), it can be shown

$$\left\langle \frac{1}{s} Za_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma}, Y - \epsilon - \frac{1}{s} Za_0^\flat \right\rangle \leq \left\| Y - \epsilon - \frac{1}{s} Za_0^\flat \right\|_2 \left\| \frac{1}{s} Za_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma} \right\|_2 \leq \frac{C_5}{s} \|a_{\text{con}}^\sharp 1\|_{\|a_{\text{con}}^\sharp\| \geq \gamma} \leq \frac{C_5}{s} \|a_{\text{con}}^\sharp 2\| \left\{ \sum_{r=1}^{R_0} \frac{\|\text{vec}(B_{0r})\|_1}{s} \right\} n \sqrt{h_n \over K} \leq \frac{C_5}{s} \|a_{\text{con}}^\sharp 2\| \left\{ \sum_{r=1}^{R_0} \frac{\|\text{vec}(B_{0r})\|_1}{s} \right\} n \sqrt{h_n \over K} \leq \frac{C_5}{s} \|a_{\text{con}}^\sharp 2\| \left\{ \sum_{r=1}^{R_0} \frac{\|\text{vec}(B_{0r})\|_1}{s} \right\} n \sqrt{h_n \over K}. \hspace{1cm} (A.85)$$

55
with probability at least

$$1 - C_6 \exp \left\{ -C_7 \left( R^2 \log \frac{\delta sK0}{\gamma} + R \sum_{d=1}^{D+1} p_d \right) \right\}.$$ 

Thus, on the interaction of the above two events, we obtain

$$\frac{1}{s^2} \| Z a_{\text{con}}^\sharp \|_2 \geq \gamma \| a_{\text{con}}^\sharp \|_2 \leq 2C_2 \frac{1}{s} \| a_{\text{con}}^\sharp \|_2 (nh) \frac{1}{2} \left( R^2 \log \frac{\delta sK0}{\gamma} + R \sum_{d=1}^{D+1} p_d \right)^{1/2}$$

$$+ 2C_5 \gamma \sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1 \frac{n\sqrt{h}}{K},$$

(A.82)

with probability at least

$$1 - C_8 \exp \left\{ -C_9 \left( R^2 \log \frac{\delta sK0}{\gamma} + R \sum_{d=1}^{D+1} p_d \right) \right\}. $$

Let

$$\delta_5 = C_1 \left\{ \sum_{r=1}^{R_0} \| \text{vec}(B_{0r}) \|_1 \right\} \frac{1}{K^\tau - 1/2} + C_2 K^{1/2} \frac{1}{\sqrt{n}} \left( R^2 \log \frac{\delta sK0}{\gamma} + R \sum_{d=1}^{D+1} p_d \right)$$

and combing (A.81) and (A.82) leads us to

$$\frac{2}{s^2} \| a_{\text{con}}^\sharp \|_2 \geq \gamma \| a_{\text{con}}^\sharp \|_2 \leq \frac{1}{s} \delta_5 \| a_{\text{con}}^\sharp \|_2,$$

with probability at least

$$1 - C_3 \exp \left\{ -C_4 \left( R^2 \log \frac{\delta sK0}{\gamma} + R \sum_{d=1}^{D+1} p_d \right) \right\}. $$

(A.83)

Note that

$$\| a_{\text{con}}^\sharp \|_2 = \| a_{\text{con}}^\sharp \|_2 \| a_{\text{con}}^\sharp \|_{< \gamma} + \| a_{\text{con}}^\sharp \|_2 \| a_{\text{con}}^\sharp \|_{\geq \gamma} \|_2 \leq 2\gamma^2 + 2\| a_{\text{con}}^\sharp \|_2 \| a_{\text{con}}^\sharp \|_{\geq \gamma} \|_2.$$ 

It then implies

$$\frac{1}{s^2} \| a_{\text{con}}^\sharp \|_2 \leq \frac{1}{s} \delta_5 \| a_{\text{con}}^\sharp \|_2 + \frac{2\gamma^2}{s^2},$$

and

$$\frac{1}{s} \| a_{\text{con}}^\sharp \|_2 \leq \frac{\left( \delta_5^2 + C_5 \gamma^2 / s^2 \right)^{1/2} + \delta_5}{2},$$

with probability at least as large as (A.83). Therefore,

$$\| \hat{m}_{\text{con}} - m_0 \| \leq C_6 \max \left\{ \frac{\delta_5^2}{K}, \gamma^2 / s^2 \right\},$$

with probability at least as large as (A.83), which completes the proof by noting $\gamma = 1/n$. 

56
We borrow the main framework in Suzuki (2015). However, there are many missing steps in the proof of Theorem 4 in Suzuki (2015). We use our own arguments to complete those details and obtain slightly different results. We then combine the nonparametric part to obtain the desired result. For a matrix \( B \), we use \( B_{:,r} \) to denote its \( r \)-th column.

**Lemma A.12.** There exists a set \( \mathcal{B}_d \subset \{ B \in \mathbb{R}^{p_d \times R} : \| B_{:,r} \|_2^2 = \gamma^2 \} \) satisfying that

\[
\begin{align*}
|\mathcal{B}_d| & \geq C_1^R \left( \frac{1}{\varrho} \right)^{R_{p_d}^2 \gamma^2}, \\
\| B - B' \|_2^2 & \geq C_2 \gamma^2 \varrho^2 R, \\
\| B_{:,r} - B'_{:,r} \|_2 & \leq \gamma,
\end{align*}
\]  

(A.84)

where \( R \geq 4, 0 < C_1, C_2 \leq 1, \varrho \leq C_3 \leq 1, \) and \( B, B' \in \mathcal{B}_d \) with \( B \neq B' \).

**Proof.** Let \( \tilde{B} \) be a \( \varrho \gamma \)-cover of the \( p_d \)-dimensional ball with radius \( \gamma \). It follows from Corollary 4.2.13 of Vershynin (2018) that \( |\tilde{U}_d| \geq \left( \frac{1}{\varrho} \right)^{p_d \gamma^2} \). Denote \( \tilde{U}_d = \{ (\tilde{\beta}_{1,d}, \ldots, \tilde{\beta}_{R,d}) : \tilde{\beta}_{r,d} \in \tilde{B}, r = 1, \ldots, R \} \).

We then construct a suitable packing set based on \( \tilde{U}_d \). In particular, by Lemma 4.2.8 of Vershynin (2018) and Lemma 5 of Suzuki (2015), there exists a \( \varrho \gamma \sqrt{R}/4 \)-packing set \( \tilde{\mathcal{B}}_d \) of \( \tilde{U}_d \) such that

\[
|\tilde{\mathcal{B}}_d| \geq \left( \frac{1}{\varrho} \right)^{R_{p_d}^2 / 4}.
\]  

(A.85)

Analogously, let \( \tilde{B}_{:,r} \) be the \( r \)-th column of \( B \in \tilde{B}_d \). We can orthogonally decompose \( \tilde{B}_{:,r} \) as

\[
\tilde{B}_{:,r} = u_r \mathbf{1} + b_r,
\]  

(A.86)

where \( u_r \) is a constant, \( \mathbf{1} \in \mathbb{R}^{p_d} \) is a vector with all elements equal to 1, \( b_r \in \mathbb{R}^{p_d} \) and \( \langle \mathbf{1}, b_r \rangle = 0 \). Since the norm of \( \tilde{B}_{:,r} \) is \( \gamma \), the orthogonality implies

\[
p_d u_r^2 + \| b_r \|_2^2 = \| \tilde{B}_{:,r} \|_2^2 \leq \gamma^2.
\]

Thus,

\[
u_r := \left( \frac{\gamma^2 - \| b_r \|_2^2/16}{p_d} \right)^{1/2}
\]  

(A.87)

is a well-defined positive value. After denoting

\[
c_r = -u_r / 4 + \bar{u}_r,
\]

(A.88)

we then have

\[
\left\| \frac{1}{4} \tilde{B}_{:,r} + c_r \mathbf{1} \right\|_2^2 = \left\| \frac{1}{4} b_r + \bar{u}_r \mathbf{1} \right\|_2^2 = \gamma^2.
\]

Let

\[
\hat{\mathcal{B}}_d = \left\{ B \in \mathbb{R}^{p_d \times R} : B_{:,r} = \frac{1}{4} \tilde{B}_{:,r} + c_r \mathbf{1}, \tilde{B} \in \tilde{B}_d, \| B_{:,r} \|_2^2 = \gamma^2, \right\}
\]  

(A.89)

where \( c_r \) is defined as (A.88).
We next show two facts, i.e.,

\[ \| \mathbf{b}_{r} - \mathbf{b}'_{r} \|_2 \leq \gamma, \quad (A.90) \]

and

\[ | \hat{\mathbf{d}} | \geq 9^{-R} \left( \frac{1}{\varrho} \right)^{R_p d/4 - R} \quad (A.91) \]

To show (A.90), we write \( c_r, r = 1, \ldots, R \) and the matrix \( \mathbf{B} \in \hat{\mathbf{d}} \) associated with \( \tilde{\mathbf{B}} \in \hat{\mathbf{d}} \) according to (A.88) and (A.89). The same goes for \( c'_r, r = 1, \ldots, \hat{R} \) and \( \mathbf{B}' \) (associated with \( \tilde{\mathbf{B}}' \)). Similarly, (A.86) and (A.87) lead to \( (u_r, \bar{u}_r, b_r, r = 1, \ldots, R) \) and \( (u'_r, \bar{u}'_r, b'_r, r = 1, \ldots, \hat{R}) \) associated with \( \mathbf{B} \) and \( \mathbf{B}' \), respectively. Thus, using the orthogonality of (A.86), we have

\[ \max \{ \| u_r,1/4 \|^2, \| b_r,4 \|^2 \} \leq \| u_r,1/4 \|^2 + \| b_r,4 \|^2 = \| \mathbf{b}_{r},4/4 \|^2 \leq \gamma^2/16, \quad (A.92) \]

and

\[ \| \mathbf{b}_{r},4 \| = \| b_r,4 + \bar{u}_r,1 \| = \| b_r,4 \|^2 + \| \bar{u}_r,1 \|^2 = \gamma^2. \quad (A.93) \]

Since (A.92) is also true when we replace \((u_r, b_r, \mathbf{B}_{r},r)\) with \((u'_r, b'_r, \tilde{\mathbf{B}}_{r},r)\), it implies that

\[ \| b_r,4 - b'_r,4 \|^2 \leq \gamma^2/4. \quad (A.94) \]

Using (A.92) and (A.93), we also obtain

\[ \frac{15}{16} \gamma^2 \leq \| \bar{u}_r,1 \|^2 \leq \gamma^2. \]

Due to \( \bar{u}_r \geq 0 \) and that the aforementioned displayed inequalities hold for \( \bar{u}'_r \) as well, we have

\[ \| \bar{u}_r,1 - \bar{u}'_r,1 \|^2 \leq (4 - \sqrt{15})^2 \gamma^2/16 < \gamma^2/4. \quad (A.95) \]

It thus follows from (A.94) and (A.95) that

\[
\| \mathbf{b}_{r} - \mathbf{b}'_{r} \| = \| b_r,4 - b'_r,4 + \bar{u}_r,1 - \bar{u}'_r,1 \|^2 \\
\leq 2\| b_r,4 - b'_r,4 \|^2 + 2\| \bar{u}_r,1 - \bar{u}'_r,1 \|^2 \\
\leq \gamma^2,
\]

which completes the proof of (A.90).

We next turn to the proof of (A.91). By definition, we obtain

\[ \mathbf{B} = \frac{1}{4} \tilde{\mathbf{B}} + \mathbf{c}^\top \otimes \mathbf{1} \]

and

\[ \mathbf{B}' = \frac{1}{4} \tilde{\mathbf{B}}' + (\mathbf{c}')^\top \otimes \mathbf{1}, \]

where \( \mathbf{c} = (c_1, \ldots, c_R)^\top \) and \( \mathbf{c}' = (c'_1, \ldots, c'_R)^\top \). After denoting \( \mathbf{u} = (u_1, \ldots, u_R)^\top \) and \( \mathbf{B} = (b_1, \ldots, b_R) \), (A.86) implies \( \tilde{\mathbf{B}} = \mathbf{B} + \mathbf{u}^\top \otimes \mathbf{1} \). Similar, we can write \( \mathbf{B}' = \mathbf{B}' + (\mathbf{u}')^\top \otimes \mathbf{1} \). If \( \mathbf{B} = \mathbf{B}' \), we much have \( \tilde{\mathbf{B}} = \tilde{\mathbf{B}}' \) since \( \mathbf{B} \) and \( \mathbf{B}' \) are the projections of \( \mathbf{B} \) and \( \mathbf{B}' \) onto the
orthogonal complement of 1, respectively. Due to \( \|\hat{B} - \hat{B}'\|^2 \geq \varrho^2 \gamma^2 R/4 \), it further shows that
\[
\|u^\top \otimes 1 - (u')^\top \otimes 1\|^2 \geq \varrho^2 \gamma^2 R/4,
\]
i.e.,
\[
\|u^\top - (u')^\top\|^2 \geq \varrho^2 \gamma^2 R/(4p_d).
\] (A.96)
The fact that \( \max\{\|\hat{B}\|^2, \|\hat{B}'\|^2\} \leq R\gamma^2 \) yields
\[
\max\{\|u^\top \otimes 1\|^2, \|(u')^\top \otimes 1\|^2\} \leq R\gamma^2,
\]
which implies
\[
\max\{\|u^\top\|^2, \|(u')^\top\|^2\} \leq R\gamma^2/p_d.
\] (A.97)
In other words, both (A.96) and (A.97) are the consequences of assuming \( B = B' \). By Lemma 4.2.8 and Corollary 4.2.13 of Vershynin (2018), the \( \sqrt{\varrho^2 \gamma^2 R/(4p_d)} \)-packing number of the \( R \)-dimensional Euclidean ball with radius \( \sqrt{R\gamma^2/p_d} \) is upper bounded by
\[
\left(1 + \frac{2}{\varrho/4}\right)^R \leq \left(\frac{9}{\varrho}\right)^R.
\]
Therefore, one \( B \) at most corresponds to \( (9/\varrho)^R \) different elements in \( \hat{B}_d \), which implies (A.91).

To finish the final proof, we use \( B \in \hat{B}_d \) as a core. Denote
\[
\mathcal{A}_\gamma(B) = \{B': \|B' - B\|^2 < \gamma^2 \varrho^2 R/(8 \times 16 \times 4), B' \in \hat{B}_d\}.
\]
Note that each element in \( \hat{B}_d \) will generate an element in \( \hat{B}_d \) according to (A.88) and (A.89). Thus, there exists a subset of \( \hat{B}_d \) to generate \( \mathcal{A}_\gamma(B) \). We denote this subset as \( \hat{A}_\gamma(B) \). These facts imply that
\[
|\mathcal{A}_\gamma(B)| \leq |\hat{A}_\gamma(B)|.
\]
Let
\[
\bar{A}_\gamma = \{B': \hat{B}' = \hat{B} + (u')^\top \otimes 1, B' \in \hat{A}(\hat{B})\}
\]
and
\[
\bar{U}_\gamma = \{u': \hat{B}' = \hat{B} + (u')^\top \otimes 1, B' \in \hat{A}(\hat{B})\}.
\]
The orthogonality between \( \hat{B} \) and 1 implies that
\[
\|\hat{B}_1 - B\|^2 < \gamma^2 \varrho^2 R/(8 \times 4)
\]
and
\[
\|\hat{B}_2 - \hat{B}\|^2 < \gamma^2 \varrho^2 R/(8 \times 4),
\]
for \( \hat{B}_1, \hat{B}_2 \in \bar{A}_\gamma \). Thus, we have
\[
\|\hat{B}_1 - \hat{B}_2\|^2 < \gamma^2 \varrho^2 R/8.
\]
Suppose \( \hat{B}_1 \) and \( \hat{B}_2 \) correspond to \( (u_1, B_1) \) and \( (u_2, B_2) \), respectively. \( \hat{B}_1 \neq \hat{B}_2 \) implies
\[
\|\hat{B}_1 - \hat{B}_2\|^2 = \|B_1 - B_2\|^2 + \|u_1^\top \otimes 1 - u_2^\top \otimes 1\|^2 \geq \varrho^2 \gamma^2 R/4.
\]
The above two displayed inequalities imply that \( \| u_1^2 \otimes 1 - u_2^2 \otimes 1 \|^2 \geq \gamma^2 R / 8 \), i.e.,

\[
\| u_1 - u_2 \|^2 \geq \gamma^2 R / (8p_d).
\]

The aforementioned inequality together with (A.97) shows a packing net property of \( \tilde{U}_r \).

Using Lemma 4.2.8 and Corollary 4.2.13 of Vershynin (2018) one more time, we have

\[
|\tilde{U}_r| \leq \left( 1 + \frac{8\sqrt{2}}{\varrho} \right) R \left( \frac{13}{\varrho} \right)^R.
\]

We finally conclude that the one element in \( \tilde{U}_r \) cannot be associated with two different elements in \( \tilde{A}_\gamma \) by showing a contradiction to the squared distance lower bound \((\gamma^2 R / 4)\) if so. Therefore, we have

\[
|\tilde{A}_\gamma(B)| \leq |\tilde{A}_\gamma(B)| \leq \left( \frac{13}{\varrho} \right)^R.
\]

(A.98)

Using (A.90), (A.91) and (A.98), we can obtain a set \( B_d \subset \tilde{B}_d \) by removing \( \tilde{A}_\gamma(B) \) from \( \tilde{B}_d \) for each \( B \in \tilde{B}_d \), which satisfies (A.84) and thus completes the proof.

\[ \square \]

Lemma A.13. There exists a subset \( W \subset \{ W \in \mathbb{R}^{M \times R}, \| W \|_2 = M \} \) such that

\[
|W| \geq C_1 R \left( \frac{1}{\varrho} \right)^{RM / 4 - 2R},
\]

\[
\| W - W' \|_2 \geq C_2 \gamma^2 M R,
\]

and

\[
\| W_{:,r} - W'_{:,r} \|_2 \leq M,
\]

where \( R \geq 4, 0 < C_1, C_2 < 1, 0 < \varrho \leq C_3 < 1 \), and \( W, W' \in W \) with \( W \neq W' \).

Proof. Taking \( \gamma = \sqrt{M} \) in Lemma A.12 will finish the proof.

One condition in Lemmas A.12 and A.13 is \( R \geq 4 \), which is also used in Lemma 5 of Suzuki (2015). Indeed, we are able to relax the condition and a slightly weaker result can be obtained below as the price paid for a milder assumption. Before presenting Lemma A.14, we let \( \tilde{B} \subset \mathbb{R}^{p_d} \) denote a \( \gamma \)-packing set of the ball with radius \( \gamma \) (\( 0 < \varrho < 1 \) without loss of generality). When a \( \gamma \)-packing set \( \tilde{B} \) of the \( p_d \)-dimensional ball with radius \( \gamma \) is given, we also define a \( \tilde{U}_d \) as the Cartesian product of \( \tilde{B} \), i.e.,

\[
\tilde{U}_d = \{ (\beta_{1,d}, \ldots, \beta_{R,d}) : \beta_{r,d} \in \tilde{B}, r = 1, \ldots, R \} \subset \{ B \in \mathbb{R}^{p_d \times R} : \| B_{:,r} \|_2 \leq \gamma^2 \}. \quad \text{(A.99)}
\]

Lemma A.14. \( \tilde{U}_d \) defined in (A.99) is a \( \gamma \)-packing set of the Cartesian product of \( p_d \)-dimensional ball with radius \( \gamma \) induced by \( \tilde{B} \). Further, there exists \( \tilde{B} \) such that \( |\tilde{U}_d| \geq (1/\varrho)^{R p_d} \).

Proof. To show \( \tilde{U}_d \) is a \( \gamma \)-packing set induced by \( \tilde{B} \), we observe that when \( B, B' \in \tilde{U}_d \) and \( B \neq B' \), there exists an \( r \in \{ 1, \ldots, R \} \) such that \( B_{:,r} \neq B'_{:,r} \). It thus implies that \( \| B - B' \|^2 \geq \| B_{:,r} - B'_{:,r} \|^2 \geq \varrho^2 \gamma^2 \). For the second statement, due to Lemma 4.2.8 and Corollary 4.2.13 of Vershynin (2018), there is a \( \gamma \)-packing set \( \tilde{B} \) of the ball in \( \mathbb{R}^{p_d} \) with radius \( \gamma \) such that \( |\tilde{B}| \geq (1/\varrho)^{p_d} \). Using the definition of \( \tilde{U}_d \) in (A.99), \( |\tilde{U}_d| \geq (1/\varrho)^{R p_d} \) is directly obtained through the Cartesian product of \( R \) copies.

\[ \square \]
A.7 Discussion on an improved rate

In this subsection, we present some discussions on the upper bound in Theorem 1. Note that by introducing a penalty that satisfies Assumption 1, \( R^{D+1} \) in the right hand side of (17) may be reduced to \( R^2 \log \delta_{\text{pen}} \), where \( \log \delta_{\text{pen}} \) is upper bounded by \( C \log \{ \max \{ n, 1 / \bar{\omega}(n), \beta_{0r,l}, \nu_0, \alpha_{0r,k} \} \} \). Roughly speaking, \( \bar{\omega}(n) \) is positively related to the penalty function. If the value of penalty term is relatively small, then the introduced bias \( G_0/n \) is dominated by the previous two terms in the right hand of (17). On the hand, both \( G_0 \) and \( 1/\bar{\omega}(n) \) are influenced by the magnitude of the true CP parameters \( \beta_{0r,l} \)'s and approximated spline coefficients \( \alpha_{0r,k} \)’s. Lemma A.15 below demonstrates the upper bound of \( G_0 \) for two cases: one is for general case and the other is for a case fulfilling Assumption S.6.

**Assumption S.6.** Suppose \( \{ \tilde{b}_k^\circ \}_{k=1}^K \) is an equivalent basis to the truncated power basis and \( \tilde{b}_1^\circ(x) = 1 \). For each \( r = 1, \ldots, R_0 \), assume there exists \( \alpha_{0r}^\circ = (\alpha_{0r,1}^\circ, \ldots, \alpha_{0r,K}^\circ)^\top \), such that

\[
\left\| f_{or} - \sum_{k=1}^K \alpha_{0r,k}^\circ \tilde{b}_k^\circ \right\|_\infty = O(K^{-\tau}),
\]

and

\[
\| \alpha_{0r}^\circ \|^2 \leq K^C.
\]

Accordingly, for any basis satisfied Assumption S.6, the upper bound of (17) in Corollary 2 can be sharpened as

\[
\| \hat{m}_{\text{PLS}} - m_0 \|^2 \leq C_4 \left\{ \min(R^{D+1}, R^2 \log n) + \frac{\sum_{i=1}^D R p_i + R K}{n} \right\} + C_5 \frac{R^2}{K^{2\tau}}. \tag{A.100}
\]

To be more specific, for any basis satisfied Assumption S.6, by Lemma A.15, we have \( G_0 \leq C_1 \lambda_1 n^{C_2} \), which yields

\[
\delta_B \leq \frac{C_1 R_2^3 + C_2 n}{\lambda_1} + C_3 n^{C_4}.
\]

Taking

\[
\frac{C_5}{C_1 n^{C_2}} \leq \lambda_1 \leq C_6 \frac{R^2 + \sum d p_d R + K R}{C_1 n^{C_2}},
\]

we obtain

\[
G_0 \leq C \left( R^2 + \sum_d p_d R + K R \right),
\]

and

\[
\delta_B \leq C_1 n^{C_2}.
\]

We then have \( \delta_{\text{pen}} \leq C_1 n^{C_2} \) and \( \log \delta_{\text{pen}} \leq C_1 \log n \), which yields (A.100).

In our implement, we used the truncated power basis \( \{ \tilde{b}_k(x) \}_{k=2}^K \) and whether this basis directly satisfying the second condition in Assumption S.6 is unknown. However, we can make an orthonormal transformation on \( \{ \tilde{b}_k(x) \}_{k=2}^K \) to obtain \( \{ \tilde{b}_k^\circ(x) \}_{k=1}^K \) to make Assumption S.6 be fulfilled under mild condition. Indeed, let \( Q^{\circ} \in \mathbb{R}^{(K-1) \times (K-1)} \) denote integrated gram matrix of \( \{ \tilde{b}_k(x) \}_{k=2}^K \), i.e., \( Q^{\circ} \) is the lower \( (K-1) \times (K-1) \) major submatrix of \( \int_{[0,1]} b(x) \{ b(x) \}^\top dx \). Since \( \{ \tilde{b}_k(x) \}_{k=1}^K \) is equivalent to B-spline with the same order and
knots (Ruppert et al., 2003), the condition in Assumption S.6 is satisfied. Moreover, it also reveals that \( Q^o \) is positive definite (for a fixed \( K \)). Therefore, we can construct \( \{b_k^o(x)\}_{k=1}^K \) through

\[
b_1^o(x) = 1 \quad \text{and} \quad \text{vec}\{b_2^o(x), \ldots, b_K^o(x)\} = (Q^o)^{-1/2} \cdot \text{vec}\{\tilde{b}_2(x), \ldots, \tilde{b}_K(x)\}. 
\]

It is not hard to see that when the magnitude of \( f_{0r} \) is upper bounded, the corresponding coefficients of \( b^o(x) \) basis are upper bounded due to their orthonormality and Lemma A.6.

Although we take the elastic-net penalty (12) with (14) and \( \lambda_2 = 0 \) in Corollary 2, the upper bound also holds for \( \lambda_2 \in [0, 1] \). Note that the proof of Corollary 2 is based on Lemmas A.10 and A.15, which also work for \( \lambda_2 \in [0, 1] \). Using similar arguments, Corollary 2 can be generalized to a version that allows \( \lambda_2 \in [0, 1] \).

As for comparing with the minimax lower bound, if we use a basis satisfied Assumption S.6 in the optimization (10), then (23) can be replaced by

\[
\max \left[ \left( R_0 n \right)^{\frac{1}{2(r+1)}}, \min \{ R_0^{D}, R_0 \log n \} \right],
\]

which is milder than the requirement of \( \sum_d p_d \) used in (23).

**Lemma A.15.** Suppose \( G(\theta) \) is defined as in (12) with (14) and \( \lambda_1 > 0 \). Under Assumption 5, if the true function \( m_0 \in \mathcal{M}_{00} \), then there exists a \( \tilde{A}_0 \) in (A.8) such that

\[
G_0 \leq C_1 \lambda_1 R_0 (K/C_2)^{2K/D} \sum_{d=1}^D p_d^2,
\]

where \( \mathcal{M}_{00} \) and \( G_0 \) are defined in (20) and (A.9), respectively. Further, if Assumption S.6 holds, then

\[
G_0 \leq C_3 \lambda_1 R_0 K C_4 \sum_{d=1}^D p_d^2
\]

**Proof.** By definition, for each \( r = 1, \ldots, R \),

\[
\int f_{0r}^2(x) dx \leq C.
\]

By Lemma A.6, there exists \( \alpha_{0r,k}, k = 1, \ldots, K \), such that

\[
\left\| f_r - \sum_{k=1}^K \alpha_{0r,k} b_k \right\|_\infty = \mathcal{O}(K^{-\tau}),
\]

which yields

\[
\int \left\{ \sum_{k=1}^K \alpha_{0r,k} b_k(x) \right\}^2 dx = \int \left\{ \sum_{k=1}^K \alpha_{0r,k} b_k(x) - f_r(x) + f_r(x) \right\}^2 dx 
\leq 2 \int \left\{ \sum_{k=1}^K \alpha_{0r,k} b_k(x) - f_r(x) \right\}^2 dx + 2 \int f_{0r}^2(x) dx 
\leq C_1 + \frac{C_2}{K^{2\tau}}.
\]
Thus, there exists \( \alpha'_{0r} = (\alpha'_{0r,1}, \ldots, \alpha'_{0r,K})^\top \), such that

\[
\int \left\{ \sum_{k=1}^{K} \alpha'_{0r,k} \tilde{b}_k(x) \right\}^2 \, dx = \int \left\{ \sum_{k=1}^{K} \alpha_{0r,k} b_k(x) \right\}^2 \, dx \leq C_1 + \frac{C_2}{K^{2\tau}} \leq C_3.
\]

Define a matrix \( \tilde{Q} \in \mathbb{R}^{K \times K} \) such that

\[
\tilde{Q}_{k_1,k_2} = \int \tilde{b}_{k_1}(x) \tilde{b}_{k_2}(x) \, dx.
\]

Note that \( \{ \tilde{b}_k(x) \} \) is a basis. Thus, \( \tilde{Q} \) is nonsingular and symmetric. Recall the truncated power basis,

\[
\tilde{b}_1(x) = 1, \quad \tilde{b}_2(x) = x, \ldots, \tilde{b}_\zeta(x) = x^{\zeta-1},
\]

\[
\tilde{b}_{\zeta+1}(x) = (x - \xi_2)^{\zeta-1}, \ldots, \tilde{b}_K(x) = (x - \xi_K-\zeta+1)^{\zeta-1}.
\]

We then have

\[
\lambda_{\max}(\tilde{Q}) \geq \frac{1}{K} \text{tr}(\tilde{Q}) \geq \frac{1}{K} \tilde{Q}_{1,1} = \frac{1}{K},
\]

where \( \lambda_{\max}(\tilde{Q}) \) denote the maximum eigenvalue of \( \tilde{Q} \). We want to find a lower bound of the minimum eigenvalue of \( \tilde{Q} \), denoted by \( \lambda_{\min}(\tilde{Q}) \).

To obtain the lower bound of \( \lambda_{\min}(\tilde{Q}) \), we consider two cases, i.e.,

\[
\lambda_{\max}(\tilde{Q}) \geq \lambda_{\min}(\tilde{Q}) \geq \frac{1}{2} \lambda_{\max}(\tilde{Q}),
\]

and

\[
\lambda_{\min}(\tilde{Q}) < \frac{1}{2} \lambda_{\max}(\tilde{Q}).
\]

When (A.103) holds, it implies

\[
\lambda_{\min}(\tilde{Q}) \geq \frac{1}{2K}.
\]

On the other hand, (A.104) together with Theorem 3.1 of Hlavácková-Schindler (2010) leads to

\[
\lambda_{\min}(\tilde{Q}) \geq \left( \frac{\sum_k \lambda_k^2(\tilde{Q}) - K \lambda_{\max}^2(\tilde{Q})}{K[1 - \lambda_{\max}^2(\tilde{Q})/\{ \prod_k \lambda_k^2(\tilde{Q})\}^{1/K}]} \right)^{1/2}
\]

\[
\geq \left\{ \frac{C \lambda_{\max}^2(\tilde{Q})}{K \lambda_{\max}^2(\tilde{Q})/\{ \prod_k \lambda_k^2(\tilde{Q})\}^{1/2}} \right\}^{1/2}
\]

\[
\geq C \left( \prod_k \lambda_k(\tilde{Q}) \right)^{1/K} K^{1/2}
\]

\[
\geq C \frac{\lambda_{\min}^{K-1}(\tilde{Q}) \lambda_{\max}(\tilde{Q})^{1/K}}{K^{1/2}}
\]

where \( \{ \lambda_k(\tilde{Q}), k = 1, \ldots, K \} \) denote all the eigenvalues of \( \tilde{Q} \). Using (A.105) and (A.106) yields

\[
\lambda_{\min}(\tilde{Q}) \geq C_1 \min \left\{ \frac{C_2^K}{K^{K/2+1}}, \frac{1}{K} \right\}.
\]
It follows that $\|\tilde{\alpha}_0\|^2 \leq C_1(K/C_2)^K$. By definition, if $\lambda_2 < 1$, then we have

$$G_0 \leq C\lambda_1 \sum_{r=1}^{R_0} (\|\tilde{\alpha}_0\|^{2/D}\|\beta_{0r,d}\|^2 + \ldots + \|\tilde{\alpha}_0\|^{2/D}\|\beta_{0r,D}\|^2)$$

$$\leq C\lambda_1 R_0 (K/C_2)^{2K/D} \sum_{d=1}^D p_d^2.$$  

Similarly, if $\lambda_2 = 1$, it shows that

$$G_0 \leq C\lambda_1 \sum_{r=1}^{R_0} (\|\tilde{\alpha}_0\|^{2/D}\|\beta_{0r,d}\|_1 + \ldots + \|\tilde{\alpha}_0\|^{2/D}\|\beta_{0r,D}\|_1)$$

$$\leq C\lambda_1 R_0 (K/C_2)^{2K/D} \sum_{d=1}^D p_d$$

$$\leq C\lambda_1 R_0 (K/C_2)^{2K/D} \sum_{d=1}^D p_d^2,$$

which finishes the proof of (A.101). (A.102) can be proved by similar arguments. \hfill \qed

B   Algorithmic analysis

B.1 Equivalent basis

To begin with, we define or recall some notations which will be used later. Recall that $\{\tilde{b}_k(x)\}_{k=1}^K$ is the truncated power basis and $\{b_k(x)\}_{k=1}^K$ is the B-spline basis. Let $u_k = \int_0^1 b_k(x)dx$ and $\tilde{u}_k = \int_0^1 \tilde{b}_k(x)dx$. Denote $\Phi(X), \tilde{\Phi}(X) \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D \times K}$ be the tensor formed from the bases, which means $(\Phi(X))_{j,k} = b_k(X_j)$ and $(\tilde{\Phi}(X))_{j,k} = \tilde{b}_k(X_j), j \in J, k = 1, \ldots, K$. We define two function classes,

$$\mathcal{M}_1 = \left\{ m(X) : m(X) = \nu_1 + \frac{1}{s} \sum_{r=1}^R \langle \beta_{1r,1} \circ \beta_{1r,2} \circ \ldots \circ \beta_{1r,D} \circ \alpha_{1r}, \Phi(X) \rangle, \sum_{k=1}^K \alpha_{1r,k} u_k = 0 \right\},$$

and

$$\mathcal{M}_2 = \left\{ m(X) : m(X) = \nu_2 + \frac{1}{s} \sum_{r=1}^R \langle \beta_{2r,1} \circ \beta_{2r,2} \circ \ldots \circ \beta_{2r,D} \circ \alpha_{2r}, \tilde{\Phi}(X) \rangle, \sum_{k=1}^K \alpha_{2r,k} \tilde{u}_k = 0 \right\},$$

where $\nu_1 \in \mathbb{R}, \beta_{1r,d} \in \mathbb{R}^{p_d}$ and $\alpha_{1r} = (\alpha_{1r,1}, \ldots, \alpha_{1r,K})^T \in \mathbb{R}^K, r = 1, \ldots, R, d = 1, \ldots, D, l = 1, 2$. Recall that $\tilde{\Phi}(X) \in \mathbb{R}^{p_1 \times \ldots \times p_D \times K-1}$ is defined by $(\tilde{\Phi}(X))_{j,k} = \tilde{b}_{k+1}(X_j), j \in J, k = 1, \ldots, K - 1$. We define the following function class that the linear constraints are removed, i.e.,

$$\mathcal{M}_3 = \left\{ m(X) : m(X) = \nu + \frac{1}{s} \sum_{r=1}^R \langle \beta_{3r,1} \circ \beta_{3r,2} \circ \ldots \circ \beta_{3r,D} \circ \alpha_{3r}, \tilde{\Phi}(X) \rangle \right\},$$
where \( \nu_3 \in \mathbb{R} \), \( \beta_{3r,d} \in \mathbb{R}^{p_d} \), and \( \alpha_{3r} = (\alpha_{3r,1}, \ldots, \alpha_{3r,K-1})^\top \in \mathbb{R}^{K-1} \), \( r = 1, \ldots, R \), \( d = 1, \ldots, D \).

By the following Lemma B.1, we can remove the linear constraints in (7) and use any equivalent spline basis to develop our theory.

**Lemma B.1.** \( \mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}_3. \)

**Proof.** First, we will prove \( \mathcal{M}_1 = \mathcal{M}_2 \). For each \( m_1(X) \in \mathcal{M}_1 \). By the property of spline basis (see, e.g., Chapter 3 of Ruppert et al., 2003), there exists an invertible matrix \( Q \) such that \( b(x) = Qb(x) \), where \( b(x) = (b_1(x), \ldots, b_K(x))^\top \) and \( \tilde{b}(x) = (\tilde{b}_1(x), \ldots, \tilde{b}_K(x))^\top \). It is straightforward to see \( \mathcal{M}_1 = \mathcal{M}_2 \).

Secondly, we will prove \( \mathcal{M}_2 \subset \mathcal{M}_3 \subset \mathcal{M}_2 \). For notational simplicity, denote

\[
B_{lr} = \beta_{l1} \circ \ldots \circ \beta_{ld}, \quad \text{for } l = 2, 3,
\]

and \( J \in \mathbb{R}^{p_1 \times \cdots \times p_D} \) as the tensor of which all the entries are 1. For each \( m_2(X) \in \mathcal{M}_2 \), take \( B_{3r} = B_{2r}, v_3 = v_2 + 1/\sum_{r=1}^R (B_{2r}, \alpha_{2r,1}J) \) and \( \alpha_{3r,k} = \alpha_{2r,k+1} \), for \( k = 1, \ldots, K - 1 \). Then we have \( m_2(X) = m_3(X) \in \mathcal{M}_3 \) and \( \mathcal{M}_2 \subset \mathcal{M}_3 \). For each \( m_3(X) \in \mathcal{M}_3 \). Suppose \( \sum_{k=1}^{K-1} \alpha_{3r,k} \tilde{u}_{k+1} = C_r \), it is trivial to see \( \tilde{u}_1 \neq 0 \). We can choose \( \alpha_{2r,1} = -C_r/\tilde{u}_1 \), \( \alpha_{2r,k+1} = \alpha_{3r,k} \) for \( k = 1, \ldots, K - 1 \) so that \( \alpha_{2r} \) satisfies the constraint in \( \mathcal{M}_2 \). Taking \( \nu_2 = \nu_3 + \sum_{r=1}^R (B_{3r}, C_r/\tilde{u}_1 J), B_{2r} = B_{3r}, \) it is trivial to see \( m_3(X) = m_2(X) \in \mathcal{M}_2 \). Thus \( \mathcal{M}_3 \subset \mathcal{M}_2 \) and we get \( \mathcal{M}_3 = \mathcal{M}_2 \).

\[\square\]

### B.2 Rescaling strategy

For a general penalty function \( G(\cdot) \) in the additive form (12), we let \( \tilde{\rho}_{r,d} = \log \rho_{r,d} \) and the rescaling strategy (13) can be written as

\[
\arg\min_{\tilde{\rho}_{r,d}, d \in \{d : \|\beta_{r,d}\|_2 \neq 0\}} \sum_{d=1}^D P_d(\exp(\tilde{\rho}_{1,d})\beta_{1,d}, \ldots, \exp(\tilde{\rho}_{R,d})\beta_{R,d})
\]

s.t. \( \sum_{d=1}^D \tilde{\rho}_{r,d} = 0 \) and \( \tilde{\rho}_{r,d} = 0 \) if \( \beta_{r,d} = 0, \ r = 1, \ldots, R, \ d = 1, \ldots, D \).

In general, (B.1) does not have a closed form solution, but we can use the Lagrangian and Newton’s methods to solve (B.1). In particular, when the elastic-net penalty (14) is used in (12), the optimization problem (B.1) analogously becomes

\[
\arg\min_{\tilde{\rho}_{r,d}, d \in \{d : \|\beta_{r,d}\|_2 \neq 0\}} \sum_{d=1}^D \frac{1}{2}(1 - \lambda_2)\|\beta_{r,d}\|_2^2 \exp^2(\tilde{\rho}_{r,d}) + \lambda_2 \|\beta_{r,d}\|_1 \exp(\tilde{\rho}_{r,d})
\]

s.t. \( \sum_{d=1}^D \tilde{\rho}_{r,d} = 0 \) and \( \tilde{\rho}_{r,d} = 0 \) if \( \beta_{r,d} = 0, \ r = 1, \ldots, R, \ d = 1, \ldots, D \),

which is a convex problem. As for the special boundary cases of \( \lambda_2 \) (i.e., \( \lambda_2 \in \{0, 1\} \)), the closed form solutions as presented in the main text can be obtained by directly solving (B.2).
B.3 Proof of Proposition 1

By definition, for any \( G(\theta) \), a solution \( \hat{\theta} \) of (B.1) satisfies

\[
LG(\hat{\theta}) \leq LG(\theta^o).
\]

For the elastic-net penalty, (B.2) is a strictly convex problem when \( \beta_{r,d} \neq 0 \) for \( r = 1, \ldots, R, d = 1, \ldots, D \). Thus \( \hat{\theta} \) is the unique minimizer in \( \Theta(\theta) \) and we complete the proof.

B.4 Proof of Proposition 2

Due to the restriction that the norm of each column of \( \tilde{B}_{D+1} \) is 1, it can be shown that \( \tilde{B}_{D+1} \) is bounded. Letting the derivative for \( \tilde{\nu} \) be 0 leads to the profile solution

\[
\tilde{\nu} = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \frac{1}{s} \sum_{r=1}^{R} \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \alpha_r, \Phi(X_i) \rangle \right)^2. \tag{B.3}
\]

By Assumption 1, \( \theta \) and \( \tilde{\nu} \) are bounded due to

\[
\{ (\tilde{\nu}, \theta, \tilde{B}_{D+1}) : LG(\tilde{\nu}, \theta, \tilde{B}_{D+1}) \leq LG(\tilde{\nu}^{(0)}, \theta^{(0)}, \tilde{B}_{D+1}^{(0)}) \}.
\]

Therefore, \( (\tilde{\nu}^{(t)}, \theta^{(t)}, \tilde{B}_{D+1}^{(t)}) \) is a bounded sequence and there exists at least one convergent sub-sequence. Suppose \( (\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}) \) is a convergent sub-sequence and denote

\[
\lim_{t \to \infty} (\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}) = (\tilde{\nu}^{*}, \theta^{*}, \tilde{B}_{D+1}^{*}).
\]

By the continuity of \( LG(\cdot) \), we have

\[
\lim_{t \to \infty} LG(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}) = LG(\tilde{\nu}^{*}, \theta^{*}, \tilde{B}_{D+1}^{*}).
\]

Note that the constraints on \( \tilde{B}_{D+1} \) forms an oblique manifold (Absil et al., 2008)

\[
\mathcal{OB} := \{ B \in \mathbb{R}^{K \times R} : \text{diag}(B^\top B) = I_R \},
\]

where \( I_R \in \mathbb{R}^{R \times R} \) is an identity matrix. Since \( \tilde{B}_{D+1}^{(n_{t})} \) minimizes the block \( LG(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \cdot) \) due to Algorithm 1, we have

\[
\text{grad}\{LG(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})})\} := P_{B^{(n_{t})}} \{ \partial LG(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t}+1)}, \tilde{B}_{D+1}^{(n_{t})})/\partial \tilde{B}_{D+1} \} = 0,
\]

where \( P_{B^{(n_{t})}}(\cdot) \) is the projection onto the tangent space of \( \mathcal{OB} \) at \( \tilde{B}_{D+1}^{*} \). The continuity of \( \text{grad}(\cdot) \) shows that (Selvan et al., 2012)

\[
\text{grad}\{LG(\tilde{\nu}^{*}, \theta^{*}, \tilde{B}_{D+1}^{*})\} = 0. \tag{B.4}
\]

Denote the algorithmic map to for the block updating of \( \tilde{\nu} \) and \( \theta \) as \( M_{\nu}(\cdot) \) and \( M_{\theta}(\cdot) \), respectively. By (B.3), we know \( M_{\nu}(\cdot) \) is continuous. According to the iterative steps of Algorithm 1, it can be shown that

\[
LG(\tilde{\nu}^{(n_{t}+1)}, \theta^{(n_{t}+1)}, \tilde{B}_{D+1}^{(n_{t}+1)}) \leq LG(M_{\nu}(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}), \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}) \leq LG(\tilde{\nu}^{(n_{t})}, \theta^{(n_{t})}, \tilde{B}_{D+1}^{(n_{t})}),
\]

66
which yields
\[ \text{LGD}(\nu^*, \theta^*, \bar{B}_{D+1}) = \text{LGD}(\nu^*, \theta^*, \bar{B}_{D+1}). \]

Thus, we have
\[ \partial \text{LGD}(\nu^*, \theta^*, \bar{B}_{D+1})/\partial(\nu) = 0. \] (B.5)

We next show that \( M_\theta \) is continuous. Since \( M_\theta(\cdot) \) is a composition of \( D \) block updating, we only need to show the \( d \)-th updating map \( M_d(\cdot) \) for the \( d \)-th mode is continuous. For notational simplicity, we let
\[ \text{LGD}(\nu, \theta, \bar{B}_{D+1}) = \text{LGD}(\nu, \theta, \bar{B}_{D+1}), \]
where \( \nu = (\nu, \bar{B}_{-d}) \) and \( \bar{B}_{-d} = B_1 \odot \cdots \odot B_{d-1} \odot B_{d+1} \odot \cdots \odot \bar{B}_{D+1} \). The \( d \)-th block updating can be represented as
\[ M_d(\nu, \theta, \bar{B}_{D+1}) = (B_1, \ldots, B_{d-1}, \bar{M}_d(\nu), B_{d+1}, \ldots, \bar{B}_{D+1}), \]
where \( \bar{M}_d(\cdot) \) is the corresponding map of updating. To show \( M_\theta \) is continuous, we thus only need to show \( \bar{M}_d(\cdot) \) is continuous for each \( d \). In the following proof, we focus on a fixed \( d \). Suppose \( \bar{z}^i \) is any sequence that converges to \( \bar{z} \). We only need to show
\[ \lim_{i \to \infty} \bar{M}_d(\bar{z}^i) = \bar{M}_d(\bar{z}). \] (B.6)

By definition and Assumption 1, \( \bar{M}_d(\bar{z}^i) \) is bounded and hence a convergent sub-sequence exists. Suppose \( \bar{M}_d(\bar{z}^i) \) is any convergent sub-sequence and denote \( \lim_{i \to \infty} \bar{M}_d(\bar{z}^i) = M_{\text{tem}} \). By the continuity of \( \text{LGD}(\cdot) \), we have
\[ \lim_{i \to \infty} \text{LGD}(\bar{z}^i, \bar{M}_d(\bar{z}^i)) = \text{LGD}(\bar{z}, M_{\text{tem}}), \]
\[ \lim_{i \to \infty} \text{LGD}(\bar{z}, \bar{M}_d(\bar{z}^i)) = \text{LGD}(\bar{z}, M_{\text{tem}}), \]
and
\[ \lim_{i \to \infty} \text{LGD}(\bar{z}^i, \bar{M}_d(\bar{z})) = \text{LGD}(\bar{z}, \bar{M}_d(\bar{z})). \]

Thus, for any \( \Xi > 0 \), when \( i \) is big enough, we have
\[ \text{LGD}(\bar{z}^i, \bar{M}_d(\bar{z})) - \Xi \leq \text{LGD}(\bar{z}^i, \bar{M}_d(\bar{z})) - \Xi \leq \text{LGD}(\bar{z}, \bar{M}_d(\bar{z})) \leq \text{LGD}(\bar{z}, \bar{M}_d(\bar{z}^i)), \]
which yields
\[ \text{LGD}(\bar{z}, \bar{M}_d(\bar{z})) = \text{LGD}(\bar{z}, M_{\text{tem}}). \]

The strict convexity of the penalty function and the optimality of \( M_{\text{tem}} \) imply that \( \bar{M}_d(\bar{z}) = M_{\text{tem}} \). We then have \( \lim_{i \to \infty} \bar{M}_d(\bar{z}^i) = \bar{M}_d(\bar{z}) \), which shows (B.6). Thus, \( \bar{M}_d(\cdot) \) is continuous and so is \( M_\theta(\cdot) \). The iterative steps of Algorithm 1 shows that
\[
\begin{align*}
\text{LG}(\bar{\nu}^{(n_{i+1})}, \theta^{(n_{i+1})}, \bar{B}^{(n_{i+1})}_{D+1}) &\leq \text{LG}(M_\nu(\bar{\nu}^{(n_i)}, \theta^{(n_i)}, \bar{B}^{(n_i)}_{D+1}), M_\theta(M_\nu(\bar{\nu}^{(n_i)}, \theta^{(n_i)}, \theta^{(n_i)}, \bar{B}^{(n_i)}_{D+1})), \\
&\leq \text{LG}(\bar{\nu}^{(n_i)}, \theta^{(n_i)}, \bar{B}^{(n_i)}_{D+1})),
\end{align*}
\]
which yields
\[ LG(\tilde{\nu}^*, M_\theta(\tilde{\nu}^*, \theta^*, \tilde{B}_{D+1}^*), \tilde{B}_{D+1}^*) = LG(\tilde{\nu}^*, \theta^*, \tilde{B}_{D+1}^*). \]

Thus,
\[ 0 \in \partial_\theta LG(\tilde{\nu}^*, \theta^*, \tilde{B}_{D+1}^*), \]
where \( \partial_\theta(\cdot) \) denotes the sub-gradient. It follows from (B.4), (B.5), and (B.7) that \((\tilde{\nu}^*, \theta^*, \tilde{B}_{D+1}^*)\) is a stationary point.

Note that the objective value \( LG(\tilde{\nu}^{(t)}, \theta^{(t)}, \tilde{B}^{(t)}_{D+1}) \) is bounded and monotonic. We then have
\[ \lim_{t \to \infty} LG(\tilde{\nu}^{(t)}, \theta^{(t)}, \tilde{B}^{(t)}_{D+1}) = LG(\tilde{\nu}^*, \theta^*, \tilde{B}^*_{D+1}), \]
which finishes the proof.

C Identifiability

It is noted that our theory does not require the identifiability for each component in (3). For completeness, we discuss the following identifiable problems. To begin with, we state the uniqueness of the representation (3), which means that (3) is the only possible combination of the coefficients and functions under the minimal \( R \) components. There are three complications that result in the indeterminacy, where two of them are similar to that of CP decomposition. The first is about permutation and scaling, i.e.,

1. Permutation and scaling. Permutation means that the summation of CP components can be permuted, i.e.,
\[ m(X) = \nu + \frac{1}{s} \sum_{r \in \{1, \ldots, R\}} \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle, \]
while scaling means that for any constant \( C \neq 0 \),
\[ \left\langle C \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, \frac{1}{C} F_r(X) \right\rangle = \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle, \]
where the scale \( C \) can also shift among \( \{\beta_{r,d}\}_{d=1}^{D} \).

The second is another possible combination of functions and the corresponding coefficients that can also represent \( m(X) \) in (3), with the exception of permutation and scaling, i.e.,

2. Another possible combination. \( m(X) \) can also be represented by
\[ m(X) = \nu + \frac{1}{s} \sum_{r=1}^{R} \langle \tilde{\beta}_{r,1} \circ \tilde{\beta}_{r,2} \circ \cdots \circ \tilde{\beta}_{r,D}, \tilde{F}_r(X) \rangle. \]
This other combination is possible. For example, let
\[ \tilde{F}_1(X) = \ldots = \tilde{F}_R(X) = F_1(X) = \ldots = F_R(X), \]
Besides, the constant shift also brings the indeterminacy. To be more specific, if we say the representation is identifiable if the components are unique up to permutation and scaling. To make (3) hold, the minimal representation means that there does not exist one of the following two representations for \( m(X) \).

3. Constant shift. For a constant \( C \) and a tensor \( J \in \mathbb{R}^{p_1 \times \cdots \times p_D} \) of which all the entries are 1,

\[
\langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) - CJ \rangle = \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle + C',
\]

where \( C' \) is a constant that can shift to the intercept \( \nu \) of the model (3).

To avoid constant shift, we let \( \int_0^1 f_r(x)dx = 0 \). This setting will not affect the expressive ability of the model (3). Now, we define the identifiability rigorously.

**Definition 4** (Identifiability). Suppose \( f_r \in \mathcal{F} \), where \( \mathcal{F} = \{ f : \int_0^1 f(x)dx = 0, f \in C([0, 1]) \} \), \( r = 1, \ldots, R \) and \( \{ f_r \}_{r=1}^R \) is the minimal representation to make (3) hold. The minimal representation means that there does not exist one of the following two representations for \( m(X) \), i.e.,

i. \( m(X) = \nu + \frac{1}{s} \sum_{r=1}^R \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle \),

where \( \nu \in \mathbb{R} \), \( \beta_{r,d} \in \mathbb{R}^{p_1 \times \cdots \times p_D} \), \( (F_r(X))_{i_1, \ldots, i_D} = f_r(X_{i_1, \ldots, i_D}) \in \mathcal{F} \) and \( \tilde{R} < R \), or

ii. \( m(X) = \bar{\nu} + \frac{1}{s} \sum_{r=1}^R \langle \bar{\beta}_{r,1} \circ \bar{\beta}_{r,2} \circ \cdots \circ \bar{\beta}_{r,D}, F_r(X) \rangle \),

where \( \bar{\nu} \in \mathbb{R} \), \( \bar{\beta}_{r,d} \in \mathbb{R}^{p_1 \times \cdots \times p_D} \), \( (F_r(X))_{i_1, \ldots, i_D} = \tilde{f}_r(X_{i_1, \ldots, i_D}) \in \mathcal{F} \) and \( \text{Span}\{ \tilde{f}_r \}_{r=1}^R \subseteq \text{Span}\{ f_r \}_{r=1}^R \).

We say the representation is identifiable if the components are unique up to permutation and scaling. To be more specific, if

\[
m(X) = \nu + \frac{1}{s} \sum_{r=1}^R \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle
\]

\[
= \nu + \frac{1}{s} \sum_{r=1}^R \langle \bar{\beta}_{r,1} \circ \bar{\beta}_{r,2} \circ \cdots \circ \bar{\beta}_{r,D}, \tilde{F}_r(X) \rangle,
\]

then \( \nu = \bar{\nu} \), and \( \{ (\beta_{r,1}, \beta_{r,2}, \cdots, \beta_{r,D}, F_r(X)) \}_{r=1}^R \) and \( \{ (\bar{\beta}_{r,1}, \bar{\beta}_{r,2}, \cdots, \bar{\beta}_{r,D}, \tilde{F}_r(X)) \}_{r=1}^R \) are the same up to scaling.

So far, we have demonstrated the identifiability issues and given the definition of identifiability with respect to the representation (3). We then list some sufficient conditions to achieve the identifiability, based on the fundamental idea of the identifiability for CP decomposition. Denote

\[
B_d = (\beta_{1,d}, \ldots, \beta_{R,d}) \quad d = 1, \ldots, D,
\]
and $k_{B_d}$ the $k$-rank of $B_d$, which is defined as the maximum value $k$ such that any $k$ columns are linearly independent (Kruskal, 1977; Harshman, 1984). Then the following conditions in the two cases are sufficient to achieve the identifiability.

Case 1. Require that $\{f_r(x)\}_{r=1}^R$ is linearly independent.

i. If $\sum_{d=1}^D k_{B_d} \geq R + D$, then the decomposition (3) is unique up to permutation and scaling.

ii. If $D = 2$ and $R(R - 1) \leq p_1(p_1 - 1)p_2(p_2 - 1)/2$, then the decomposition (3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

iii. If $D = 3$ and $R(R - 1) \leq p_1p_2p_3(3p_1p_2p_3 - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, then the decomposition (3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Case 2. Not require that $\{f_r(x)\}_{r=1}^R$ is linearly independent.

iv. (General) If $\sum_{d=1}^D k_{B_d} \geq 2R + D - 1$, then the decomposition (3) is unique up to permutation and scaling.

For simplicity, we present the general condition in the following theorem. In the proof of Theorem C.3, we in fact prove all the aforementioned sufficient conditions.

**Theorem C.3 (Identifiability).** If

$$\sum_{d=1}^D k_{B_d} \geq 2R + D - 1,$$

then the representation (3) is unique up to permutation and scaling.

**Proof.** Suppose there is another representation of (3), i.e,

$$m(X) = \nu + \frac{1}{s} \sum_{r=1}^R \langle \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D}, F_r(X) \rangle$$

$$= \bar{\nu} + \frac{1}{s} \sum_{r=1}^R \langle \bar{\beta}_{r,1} \circ \bar{\beta}_{r,2} \circ \cdots \circ \bar{\beta}_{r,D}, \bar{F}_r(X) \rangle,$$

where

$$(F_r(X))_{i_1i_2\cdots i_D} = f_r(X_{i_1i_2\cdots i_D}) \quad \text{and} \quad (\bar{F}_r(X))_{i_1i_2\cdots i_D} = \bar{f}_r(X_{i_1i_2\cdots i_D}),$$

with $f_r, \bar{f}_r \in \mathcal{F}$, $r = 1, \ldots, R$. We will show $\bar{\nu} = \nu$, as well as $\beta_{r,d}$ and $\bar{\beta}_{r,d}$, $f_r$ and $\bar{f}_r$, $r = 1, \ldots, R$, $d = 1, \ldots, D$, are the same up to permutation and scaling under some conditions, respectively.

Using the definition of $\mathcal{F}$, such as $\int_0^1 f(x)dx = 0$ for $f \in \mathcal{F}$, we can obtain $\nu = \bar{\nu}$ by integration over the domain of $X$ in (C.2). In the remaining sum of inner products, we consider the following arguments. Suppose the minimal bases of the vector space

$$\text{Span}\{f_r(x), r = 1, \ldots, R\} \quad \text{and} \quad \text{Span}\{\bar{f}_r(x), r = 1, \ldots, R\}$$
\[ f_r(x) = \sum_{k^* = 1}^{K^*} \eta_{r,k^*} \psi_{k^*}(x) \quad \text{and} \quad \tilde{f}_r(x) = \sum_{k^* = 1}^{\tilde{K}^*} \tilde{\eta}_{r,k^*} \tilde{\psi}_{k^*}(x). \]

For notational convenience, we let \( \Psi(X)_{j,k^*} = \psi_{k^*}(X_j) \), \( k^* = 1, \ldots, K^* \) and \( \bar{\Psi}(X)_{j,k^*} = \bar{\psi}_{k^*}(X_j) \), \( k^* = 1, \ldots, \bar{K}^* \), where \( j \in J \). We also denote

\[ A^f = \frac{1}{s} \sum_{r=1}^{R} \beta_{r,1} \circ \beta_{r,2} \circ \cdots \circ \beta_{r,D} \circ \eta_r, \quad (C.3) \]

and

\[ \bar{A}^f = \frac{1}{s} \sum_{r=1}^{R} \bar{\beta}_{r,1} \circ \bar{\beta}_{r,2} \circ \cdots \circ \bar{\beta}_{r,D} \circ \bar{\eta}_r, \quad (C.4) \]

where \( \eta_r = (\eta_{r,1}, \ldots, \eta_{r,K})^\top \) and \( \bar{\eta}_r = (\bar{\eta}_{r,1}, \ldots, \bar{\eta}_{r,K})^\top \), for \( r = 1, \ldots, R \). Since we have shown \( \nu = \bar{\nu} \) in the previous arguments, it is trivial to see that the remaining summation of CP components in (C.2) equals, i.e.,

\[ \langle A^f, \Psi(X) \rangle = \langle \bar{A}^f, \bar{\Psi}(X) \rangle. \quad (C.5) \]

The rest of proof includes three steps. At first, we will show

\[ \text{Span}\{\psi_{k^*}(x)\}_{k^* = 1}^{K^*} = \text{Span}\{\tilde{\psi}_{k^*}(x)\}_{k^* = 1}^{\tilde{K}^*}. \quad (C.6) \]

Based on (C.6), we can chose \( \{\tilde{\psi}_{k^*}(x)\}_{k^* = 1}^{\tilde{K}^*} = \{\psi_{k^*}(x)\}_{k^* = 1}^{K^*} \) and rewrite (C.5) as

\[ \langle A^f, \Psi(X) \rangle = \langle \bar{A}^f, \bar{\Psi}(X) \rangle. \quad (C.7) \]

Secondly, we will show \( A^f = \bar{A}^f \) in (C.7). In the end, we will take the advantages of identifiable theory about CP decomposition and complete the proof.

To show (C.6), we assume there exists \( k_0 \) such that \( \tilde{\psi}_{k_0}(x) \) is linearly independent of \( \{\psi_{k^*}(x)\}_{k^* = 1}^{K^*} \). For each \( j \in J \), we take integration for other predictors over their domain, then by Lemma A.1, we get

\[ \sum_{k^* = 1}^{K^*} A^f_{j,k^*} \psi_{k^*}(X_j) - \sum_{k^* \neq k_0}^{K^*} \tilde{A}^f_{j,k^*} \tilde{\psi}_{k^*}(X_j) - \tilde{A}^f_{j,k_0} \tilde{\psi}_{k_0}(X_j) = 0, \]

for \( X_j \in [0, 1] \). Note that \( \tilde{\psi}_{k_0}(x) \) is independent of \( \{\psi_{k^*}(x)\}_{k^* = 1}^{K^*} \) and \( \{\tilde{\psi}_{k^*}(x)\}_{k^* \neq k_0} \), then \( \tilde{A}^f_{j,k_0} = 0 \), for \( j \in J \). Assume there exists \( r_0 \) such that \( \bar{\eta}_{r_0,k_0} \neq 0 \), then there exists \( \tilde{f}_{r_0} \in R \), where \( \tilde{f}_{r_0}(x) = \sum_{k^* \neq k_0} \bar{\eta}_{r_0,k^*} \bar{\psi}_{i}(x) \) and \( \text{Span}\{\tilde{f}_{r_0}\}_{r = 1}^{R} \subseteq \text{Span}\{\tilde{f}_{r}\}_{r = 1}^{R} \), such the representation (3) holds. This does not agree with the minimal representation assumption. As a result, \( \bar{\eta}_{r,k_0} = 0 \) for \( r = 1, \ldots, R \), then \( \{\tilde{f}_{r}(x)\}_{r = 1}^{R} \) can be represented by \( \{\tilde{\psi}_{k^*}(x)\}_{k^* \neq k_0} \), which leads a contradiction to that \( \{\tilde{\psi}_{k^*}(x)\}_{k^* = 1}^{\tilde{K}^*} \) is a minimal basis. Therefore (C.6) holds and \( \tilde{K}^* = K^* \).
To show $A^f = \bar{A}^f$ in (C.7), we let $A^{f,*} = A^f - \bar{A}^f$. It implies that

$$\langle A^{f,*}, \Psi(X) \rangle = 0,$$

for all $X$. Assuming $A^{f,*} \neq 0$, there exists $j_0 \in \mathcal{J}$ such that $(A^{f,*}_{j_0,1}, \ldots, A^{f,*}_{j_0,K^*}) \neq 0$. We fix $\{X_j\}_{j \neq j_0}$ at some values and denote the corresponding value

$$C_{-j_0} = \sum_{j \neq j_0} \sum_{k^* = 1}^{K^*} A^{f,*}_{j,k^*} f(k^*)(X_j).$$

It then shows that

$$\sum_{k^* = 1}^{K^*} A^{f,*}_{j_0,k^*} \psi_{k^*}(X_{j_0}) + C_{-j_0} = 0,$$

for $X_{j_0} \in [0,1]$. By integration over $X_{j_0}$ on both sides, we obtain

$$\sum_{k^* = 1}^{K^*} A^{f,*}_{j_0,k^*} w_{k^*} + C_{-j_0} = 0,$$

where $w_{k^*} = \int_0^1 \psi_{k^*}(x)dx$, $k^* = 1, \ldots, K^*$. By Lemma A.1, $\sum_{k^* = 1}^{K^*} A^{f,*}_{j_0,k^*} w_{k^*} = 0$, which implies $C_{-j_0} = 0$. Combining the independence and (C.8) yields $A^{f,*}_{j_0,k^*} = 0$ for $k^* = 1, \ldots, K^*$. Thus $A^{f,*} = 0$ and we have $A^f = \bar{A}^f$.

Since $R$ is the minimal, (C.3) is a rank decomposition of $A^f$. We can claim that if the rank decomposition of $A^f$ is unique up to permutation and scaling, then the representation (3) is unique up to scaling and permutation. To see this, we can assume the rank decomposition of $A^f$ is unique up to permutation and scaling. Thus the decomposition (C.4) and the decomposition (C.3) are the same up to permutation and scaling. Therefore the representation (3) is unique up to permutation and scaling. Now, to make the representation (3) unique up to permutation and scaling, we can use the common arguments about the uniqueness of rank decomposition. Recall that $B_d = (\beta_{1,d}, \ldots, \beta_{R,d})$, $d = 1, \ldots, D$ and the $k$-rank of a matrix $B_d$, denoted as $k_{B_d}$, is defined as the maximum value $k$ such that any $k$ columns are linearly independent. For convenience, we write $B_{D+1} := \eta = (\eta_1, \ldots, \eta_R)$ and let $k_{B_{D+1}}$ be its $k$-rank. To make the CP decomposition of $A^f$ unique, we have the following sufficient conditions

1. (General) (Sidiropoulos and Bro, 2000) The decomposition (C.3) is unique up to permutation and scaling if $\sum_{d=1}^{D+1} k_{B_d} \geq 2R + D$.

2. (De Lathauwer, 2006) When $D + 1 = 3$, $R \leq K$ and $R(R-1) \leq p_1(p_1-1)p_2(p_2-1)/2$, the decomposition (C.3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

3. (De Lathauwer, 2006) When $D + 1 = 4$, $R \leq K$ and $R(R-1) \leq p_1p_2p_3(3p_1p_2p_3 - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, the decomposition (C.3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Now we consider two cases, i.e,
Case 1. If \( \{ f_r(x) \}_{r=1}^R \) is linearly independent, then \( k_{B_{D+1}} = R \). We have the following sufficient conditions.

i. If \( \sum_{d=1}^D k_{B_d} \geq R + D \), then the decomposition (3) is unique up to permutation and scaling.

ii. If \( D = 2 \) and \( R(R - 1) \leq p_1(p_1 - 1)p_2(p_2 - 1)/2 \), then the decomposition (3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measure zero.

iii. If \( D = 3 \) and \( R(R - 1) \leq p_1p_2p_3(3p_1p_2p_3 - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4 \), then the decomposition (3) is unique up to permutation and scaling for almost all such tensors except on a set of Lebesgue measures zero.

Case 2. If we do not know whether \( \{ f_r(x) \}_{r=1}^R \) is linearly independent or not, we can also use the fact that \( k_{B_{D+1}} \geq 1 \), which yields the following general sufficient condition.

iv. (General) If \( \sum_{d=1}^D k_{B_d} \geq 2R + D - 1 \), then the decomposition (3) is unique up to permutation and scaling.

Since \( \{ f_r \}_{r=1}^R \) are allowed to be the same in the model, we present the forth sufficient condition, i.e.,

\[
\sum_{d=1}^D k_{B_d} \geq 2R + D - 1,
\]

which is also used as a condition to make the tensor linear model identifiable (Zhou et al., 2013).

\[\square\]

D Starting points

Motivated by the initial point strategy used in the MATLAB toolbox TensorReg, we propose a sequential down-sizing strategy. For the penalized tensor linear regression, TensorReg applies the unpenalized tensor linear regression on a down-sized sample first. The down-sized sample depends on a shrinkage parameter \( \vartheta = n/(CR\sum_{d=1}^D p_d) \), where \( C \) is a constant supplied by users. If \( \vartheta \leq 1 \), the down-sized sample is just the original sample \( (X_i, y_i) \); if not, \( X_i \in \mathbb{R}^{p_1 \times \ldots \times p_D} \) is down-sized to a smaller tensor of size \( \tilde{p}_1 \times \ldots \times \tilde{p}_D \), where \( \tilde{p}_d = \lfloor p_d/\vartheta \rfloor \). Secondly, transform the solution of coefficient tensor of the previous unpenalized method back to the original size, and then run the penalized algorithm.

Our sequential down-sizing procedure refers to considering a sequential down-size \( \tilde{p}_1^{(1)} \times \cdots \times \tilde{p}_D^{(1)} \times K, \tilde{p}_1^{(2)} \times \cdots \times \tilde{p}_D^{(2)} \times K, \cdots, \tilde{p}_1^{(\eta)} \times \cdots \times \tilde{p}_D^{(\eta)} \times K \) of the samples in the initial stage. Here we can choose \( \{ p_d^{(i)} \}_{i=1}^{\eta} \subset (C, p_d) \cap \mathbb{N} \) as an arithmetic sequence, where \( C \) is a constant defined by users. Firstly, we use random initial points for the unpenalized tensor regression on the down-size sample with size of \( \tilde{p}_1^{(1)} \times \cdots \times \tilde{p}_D^{(1)} \times K \), and then use the results of the unpenalized tensor regression under the size of \( \tilde{p}_1^{(i)} \times \cdots \times \tilde{p}_D^{(i)} \times K \) as the initial points (after up-size) for that of \( \tilde{p}_1^{(i+1)} \times \cdots \times \tilde{p}_D^{(i+1)} \times K \), where \( 1 \leq i < \eta \). Finally, use the results (after
up-size) under the size of $\tilde{p}_1^{(n)} \times \cdots \times \tilde{p}_D^{(n)} \times K$ as the final initial points for the penalized method.

In the grids search procedure, when $R$ and $\lambda_2$ are fixed, we apply the sequential down-sizing initial strategy to the smallest $\lambda_1$ in the grids, and then use the results as the initial points of the second smallest $\lambda_1$ in grids. Next, we use the new results as the initial points of the third smallest $\lambda_1$ and repeat this procedure for all values of $\lambda_1$.

### E Data preprocessing of monkey’s electrocorticography data

Our data preprocessing procedure is similar to Chao et al. (2010) and Shimoda et al. (2012). Firstly, the original signals were band-pass filtered from 0.3 to 499Hz and re-referenced using a common average reference montage (McFarland et al., 1997). Secondly, we use Morlet wavelet transformation to get the time-frequency representation at time $t$, where there are ten different center frequencies (20Hz, 30Hz, ..., 110 Hz) and ten time lags ($t - 900$ ms, $t - 800$ ms, ..., $t - 100$ ms, $t$). Finally, after a standardization step ($z$-score) at each frequency over the 10 time lags for each electrode, we get our input tensor of size $64 \times 10 \times 10$, such that the values of each entry lie in $I = [-2.75, 2.85]$.

### F Discussion on the global optimum

Due to the non-convexity of the objective function, it is not a trivial task to theoretically fill the gap between Proposition 2 and Theorem 1. We note that the similar theoretical gap also exists in a number of statistical work, notably the tensor linear regression (Zhou et al., 2013), though many still perform reasonably well in practice. Although we assume the global optimal as our estimator, the asymptotic theory (Section 4) indeed only relies on that (A.11) holds, i.e.,

$$
\sum_{i=1}^n \left( y_i - \frac{1}{s} \langle \tilde{A}_\text{PLS}, \Phi(X_i) \rangle \right)^2 \leq \sum_{i=1}^n \left( y_i - \frac{1}{s} \langle A_0, \Phi(X_i) \rangle \right)^2 + G_0.
$$

In other words, as long as the loss function evaluated at the estimator is small enough, then error bound will hold. We call the left hand side as (LHSloss) and the right hand side as (RHSobj) in (A.11). The result of RHSobj minus LHSloss is denoted as RHSobj - LHSloss. In our numerical experiments, we found that LHSloss always smaller than RHSobj, based on our implementation of the proposed algorithm; see Figure S.1 as an example of Case 2 in the synthetic data.

### G ADNI data

We also evaluated our proposed method, as well as the alternative TLR using elastic-net penalization with rescaling strategy (TLR-rescaled), on a publicly available data set obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI, Mueller et al., 2005) database. ADNI was initialized in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and the Food and Drug Administration (FDA). Alzheimer’s disease is an age-related neurologic disorder that causes the brain to
Figure S.1: The LHS (LHS\text{loss}) and RHS (RHS\text{obj}) of (A.11) in Case 2 of the synthetic data. The first row shows these two quantities based on 50 replications after validation, when \( n = 500 \) and \( n = 1000 \). The second row depicts the differences (RHS\text{obj} - LHS\text{loss}) accordingly.

shrink (atrophy) and brain cells to die. It is often characterized by progressive memory impairment and deterioration of cognitive functions (de la Torre, 2010). The primary goal of ADNI is uniting researchers with study data to investigate whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be used to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD, Mueller et al., 2005). See more detailed descriptions of ADNI at its website (adni.loni.ucla.edu).

The data analyzed here is a set of PET from ADNI standardized using the tool \textit{dcm2niix} (Li et al., 2016) for 774 subjects, and the preprocessed image of each subject is the last output of \textit{dcm2niix} with size \( 160 \times 160 \times 96 \). Of the 774 subjects in our study, 250 have been diagnosed AD with average age 75.5 (8.2) and 524 are cognitively normal (CN) with average age 75.0 (7.3), where the numbers in brackets are the respective standard deviations. To facilitate the computation, we implemented a downsizing procedure (Zhang et al., 2017, 2019). Similar to Reiss and Ogden (2010), we chose slices (i.e., the 51st, 52nd, and 53rd slices from the bottom) of the preprocessed image to analyze. Following Zhang et al. (2020), a downsizing step was applied to each slice using the method in the MATLAB toolbox TensorReg. Finally, we obtained the tensor covariate \( \mathbf{X}_i \in \mathbb{R}^{40 \times 40 \times 3} \).

We encoded the two classes as 1 (AD) and \(-1\) (CN) as in Huang and Pan (2003); Khosla.
et al. (2018), and fitted various models to estimate \( m(\cdot) \). The data set was randomly split into three different subsets, i.e., a training set, a validation set, and a test set, of size 495, 124, and 155 respectively. The prediction procedure on the test set was based upon the sign of the predicted value. In other words, if the output was positive, we predicted 1 (AD), while if the output is negative, we predicted \(-1\) (CN). The method of selecting the tuning parameters and the corresponding grid points are the same as we used in Section 5 for the ECoG dataset. To measure the performance of prediction, we used out-of-sample classification accuracy (i.e., 1- misclassification error) based on 10 random splittings. The average classification accuracies by TLR-rescaled and BroadcasTR are 0.834 and 0.869, respectively. Figure S.2 depicts boxplots of the classification accuracy using TLR-rescaled and BroadcasTR. It shows that the performance of the proposed BroadcasTR is superior to that of the linear model in terms of prediction. Moreover, our finding that the nonlinearity can help improve the prediction in Alzheimer’s disease is confirmed by existing literature (e.g., Lu et al., 2018).

![Boxplot of Classification Accuracy](image)

**Figure S.2:** Prediction performance on the ADNI data. The left and right boxplots are respectively the classification accuracy of BroadcasTR and TLR-rescaled based on 10 random splittings.

To further show the benefit of our proposed BroadcasTR in terms of region selection, we refitted two models to all samples with the tuning parameters corresponding to the best predicted results. Since the decision rule is the sign of \( \hat{m}(X) \), we summarized the signed empirical contribution of the \((j_1, j_2, j_3)\)-th pixel using

\[
\frac{1}{n} \sum_i \hat{m}_{j_1,j_2,j_3}(X_{i,j_1,j_2,j_3}) \mathbf{1}_{\{\hat{m}_{j_1,j_2,j_3}(X_{i,j_1,j_2,j_3}) > 0\}}
\]

and

\[
\frac{1}{n} \sum_i -\hat{m}_{j_1,j_2,j_3}(X_{i,j_1,j_2,j_3}) \mathbf{1}_{\{\hat{m}_{j_1,j_2,j_3}(X_{i,j_1,j_2,j_3}) < 0\}}
\]

to reflect the important regions, where \( \mathbf{1}_{\{\cdot\}} \) is an indicator function. Figure S.3 depicts the results of the positive and negative contributions for two methods. For the positive contribution (towards AD), BroadcasTR is more concentrated in the sub-area of the medical prefrontal cortex in the default mode network (Spreng et al., 2010) compared with TLR-rescaled. It has been confirmed that during the attention demanding cognitive tasks, the activities of prefrontal cortex is observed to become weak gradually with the difficulties of the tasks (Raichle et al., 2001). Thus, the prefrontal cortex is extremely vulnerable to neurodegeneration with Alzheimer’s disease (Salat et al., 2001). For the negative contribution (towards CN), the important regions selected by TLR-rescaled and BroadcasTR are roughly identical. For example,
Figure S.3: Region selection performance on the ADNI data. The columns correspond to the slices of the tensor covariate. The rows named “Pos-” and “Neg-” are the plots of positive and negative contributions of each entry, respectively. The rows named “-TLR” and “-BroadcasTR” correspond to the tensor linear regression with rescaling strategy and the proposed broadcasted nonparametric model, respectively.
Table S.1: Prediction performance on the synthetic data. Reported are averages of MSPE and its standard deviation (in parenthesis) based on 50 replications.

| Case | n   | GPNTE  | TVGP     | BroadcasTR |
|------|-----|--------|----------|------------|
| 2    | 500 | 35.51143 (2.532002) | 31.86366 (2.547318) | 2.440391 (0.3408013) |

both methods identify a sub-area of the corpus callosum (Davatzikos et al., 1996) and Figure S.3 reveals that the size of the corpus callosum in CN is larger than that in AD patients. This phenomenon is consistent with the scientific finding that the size of the corpus callosum in AD will be significantly reduced (Teipel et al., 2002).

H An extra simulation study

We also compared our method with existing nonlinear tensor regression models, including tensor-variate Gaussian process regression (TVGP, Zhao et al., 2014) and Gaussian process nonparametric tensor estimator (GPNTE, Kanagawa et al., 2016). We conducted the comparison using the nonlinear setting Case 2 in Subsection 5.1 of the main paper with sample size \( n = 500 \), 20% of which was used for validation. Since GPNTE requires the tensor covariate to be rank one, we generated the covariate as

\[ X = x_1 \odot x_2 \in \mathbb{R}^{64 \times 64}, \]

where each entry of \( x_d \) was independently sampled from Uniform\([0,1]\), \( d = 1, 2 \). The method of selecting the tuning parameters and the corresponding grid points are the same as we used in Section 5 for the synthetic data. After training the proposed BroadcasTR and these two alternatives, we independently generated another test dataset with 500 sample size and the mean squared prediction error (MSPE) defined in (24) was calculated for various methods to measure their performance. We report the prediction comparison in Table S.1 based on 50 replications.

Table S.1 shows that the performance of TVGP and GPNTE is inferior to the proposed BroadcasTR. The bad performance of these alternatives is attributed to that TVGP and GPNTE have to intrinsically estimate a 64^2-dimensional and multiple 64-dimensional functions, respectively. Our proposed BroadcasTR, on the other hand, can handle a relatively high-dimensional setting by broadcasting multiple unidimensional functions.