Journal of Machine Learning Research 23 (2022) 1-32 Submitted 7/18; Revised 10/21; Published 12/21

# Debiased Distributed Learning for Sparse Partial Linear Models in High Dimensions

**Shaogao Lv** LVSG716@NAU.EDU.CN
*Department of Statistics and Data Science*
*Nanjing Audit University*
*Nanjing, China*

**Heng Lian** HENGLIAN@CITYU.EDU.HK
*Department of Mathematics*
*City University of Hong Kong*
*Hong Kong, China*
*and*
*City University of Hong Kong Shenzhen Research Institute*
*Shenzhen, China*

**Editor:** Lorenzo Rosasco

## Abstract

Although various distributed machine learning schemes have been proposed recently for purely linear models and fully nonparametric models, little attention has been paid to distributed optimization for semi-parametric models with multiple structures (e.g. sparsity, linearity and nonlinearity). To address these issues, the current paper proposes a new communication-efficient distributed learning algorithm for sparse partially linear models with an increasing number of features. The proposed method is based on the classical divide and conquer strategy for handling big data and the computation on each subsample consists of a debiased estimation of the doubly regularized least squares approach. With the proposed method, we theoretically prove that our global parametric estimator can achieve the optimal parametric rate in our semi-parametric model given an appropriate partition on the total data. Specifically, the choice of data partition relies on the underlying smoothness of the nonparametric component, and it is adaptive to the sparsity parameter. Finally, some simulated experiments are carried out to illustrate the empirical performances of our debiased technique under the distributed setting.

**Keywords:** Big data, Distributed learning, High dimensions, Reproducing kernel Hilbert space (RKHS), Semi-parametric models

## 1. Introduction

Under a big-data setting, the storage and analysis of data can no longer be performed on a single machine, and in this case dividing data into many sub-samples becomes a critical procedure for any numerical algorithm to be implemented. Distributed statistical estimation and distributed optimization have received increasing attention in recent years, and a flurry of researches towards solving large-scale problems have emerged recently, such as Mcdonald et al. (2009); Zhang et al. (2013, 2015); Rosenblatt and Nadler (2016) and the references therein. In general, distributed algorithms can be classified into two families: data paral-

lelism and task parallelism. Data parallelism aims at distributing the data across different computing nodes or machines while task parallelism distributes different tasks across computing nodes. We are only concerned with data parallelism in this paper. In particular, we primarily consider the distributed estimation for partially linear models via the standard divide-and-conquer strategy. *Divide-and-conquer* is a simple and communication-efficient way for handling big data, which is commonly used in the literature of statistical learning. To be precise, the whole data is randomly allocated to $m$ machines, a local estimator is computed independently on each machine, and then the central node averages the local solutions into a global estimate.

Partially linear models (PLM) (Hardle and Liang, 2007; Heckman, 1986), as the leading example of semiparametric models, are a class of important tools for modeling complex data, which retain model interpretation and flexibility simultaneously. Given the observations $(Y_i, X_i, T_i), i = 1, ..., N$, where $Y_i$ is the response, $X_i = (x_{i1}, ..., x_{ip})' \in \mathbb{R}^p$ and $T_i = (t_{i,p+1}, ...., t_{i,p+q})' \in \mathbb{R}^q$ are vectors of covariates, partially linear models assume that

$$Y_i = X_i'\boldsymbol{\beta}^* + f^*(T_i) + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is a vector of unknown parameters for linear terms, $f^*$ is an unknown function defined on a compact subset $\mathcal{T}$ of $\mathbb{R}^q$ ($q$ is fixed), and $\epsilon_i$'s are independent standard normal variables. In the sparse setting, one often assumes that the cardinality of nonzero components of $\boldsymbol{\beta}^*$ is less than $p$, that is, $s^* := |S := \{j \in [p],\, \beta_j^* \neq 0\}| \ll p$.

There is a substantial body of work focusing on the sparse setting for PLM; see, for example, Green and Yandell (1985); Wahba (1990); Hardle and Liang (2007); Zhang et al. (2011); Lian et al. (2012); Wang et al. (2014), among others. Chen (1988) and Robinson (1988) showed that the parametric part can be estimated with parametric rates under appropriate conditions. Mammen and van de Geer (1997) proved the linear part is asymptotically normal under a more general setting. These results are asymptotic and valid in the fixed-dimensional case, where the number of variables $p$ in the linear part is less than the number of observations.

Although this paper is mainly concerned with data parallelism, which is practically useful in the $N > p$ setting, the size of each sub-sample ($n := N/m$) that is allocated to each node may be less than $p$ with a large number of nodes ($m$). So the high dimensional issue has been endowed with additional implications under the data parallelism setting. Compared to the linear models or the nonparametric additive models, the high dimensional case for studying PLM with $p > n$ is more challenging, mainly because of the correlation and interaction between the covariates in the linear part and the covariates in the nonparametric part. Under the high dimensional framework, a commonly-used approach is to construct penalized least squares estimators with a double penalty term, using a smoothness penalty to control the complexity of the nonparametric part and a shrinkage penalty on the parametric part to achieve model parsimony. To build each individual estimator before merging, we consider a doubly regularized approach with the Lasso penalty and a reproducing kernel Hilbert space (RKHS) norm penalty, given by

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p, f\in\mathcal{H}_K} \{\mathcal{L}^{(l)}(\boldsymbol{\beta}, f)\},\ \ \mathcal{L}^{(l)}(\boldsymbol{\beta}, f) = \frac{1}{2n}\sum_{i\in S_l}\big(Y_i - X_i'\boldsymbol{\beta} - f(T_i)\big)^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + (\lambda_2/2)\|f\|_K^2, \tag{2}$$

where $S_l$ is the $l$-th subsample with size $n$, and $(\lambda_1, \lambda_2)$ are two tuning parameters. Here we consider a function from a RKHS denoted by $\mathcal{H}_K$, endowed with the norm $\|\cdot\|_K$. The kernel function $K$ defined on $\mathcal{T} \times \mathcal{T}$ and $\mathcal{H}_K$ are determined by each other.

With a diverging dimension in the linear part, there is a rich literature on penalized estimation for PLM in the last decade. Xie and Huang (2009) proposed the SCAD-penalized estimators of the linear coefficients, and achieved estimation consistency and variable selection consistency for the linear and nonparametric components. Similar to (2), Ni et al. (2009) formulated a doubly regularized least squares approach, using the smoothing spline to estimate the nonparametric part and the SCAD to conduct variable selection. It is shown that the proposed procedure can be as efficient as the oracle estimator. Recently, Wang et al. (2014) proposed a new doubly penalized procedure for variable selection with respect to both linear and additive components, where the numbers of linear covariates and nonlinear components both diverge with the sample size. All these aforementioned papers focus on the case where $p$ is relatively small compared to $n$ (*e.g.* $p = o(\sqrt{n})$).

Allowing for $p \geq n$ and even $p = o(e^n)$, recent years has witnessed several related research in terms of non-asymptotic analysis for the sparse PLM. As shown in the distributed learning literature (Lee et al., 2017; Zhang et al., 2015), the optimal estimation of each local estimate is very critical to derive the optimal non-asymptotic results of the averaging estimate. Under the non-distributed setting, Müller and van de Geer (2015) theoretically analyzed the penalized estimation (2), and proved that the parametric part in PLM achieves the optimal rates of the linear models, as if the nonparametric component were known in advance. More recently, Zhu (2017) considered a two-step approach for estimation and variable selection in PLM. The first step uses nonparametric regression to obtain the partial residuals, and the second step is an $\ell_1$-penalized least squares estimator (the Lasso) to fit the partial residuals from the first step. Like Müller and van de Geer (2015), they derived optimal non-asymptotic oracle results for the linear estimator.

In this paper, we aim at proposing an efficient distributed estimation for high dimensional PLM. Although distributed estimation on linear models (Zhang et al., 2013; Lee et al., 2017; Battey et al., 2018) and fully nonparametric models (Zhang et al., 2015; Lin et al., 2017) have been well-understood, the investigation on distributed estimation for PLM is more challenging and there are very few works on this (Zhao et al., 2016; Lian et al., 2019). First, it is known that the Lasso penalty and the functional norm in (2) lead to a heavily biased estimation, and naive averaging only reduces the variance of the local estimators, but has no effect on the bias (Mcdonald et al., 2009; Wu, 2017). Moreover, in the diverging dimensional setting (i.e. $p, n \to \infty$), Rosenblatt and Nadler (2016) showed that the averaged empirical risk minimization (ERM) is suboptimal versus the centralized ERM. Therefore, debiasing is essential to improving accuracy of the averaging estimate. In addition, the significant influences of high dimensions and correlated covariates from the parametric and nonparametric components will result in additional technical difficulties.

Our main contribution to this line of research consists of two aspects. Our first contribution is to analyze the doubly regularized least squares method (2) for estimating sparse PLM and provide upper bounds on the parametric part and the nonparametric part respectively. These derived results based on any given subsample serve further our proposed averaging estimation by merging the total data. Our proof for optimal rates of the parameter estimator in PLM is partially inspired by the idea from Müller and van de Geer (2015),

but there exist some technical distinctions from the main proof in Müller and van de Geer (2015) under the non-distributed setting. In particular, the proof of learning rate of the parametric estimator is based on the loss function directly, while the corresponding proof in Theorem 2 of Müller and van de Geer (2015) uses the first-order condition. The latter may exclude the Lipschitz-loss based learning, such as the classical SVM and the quantile regression. It is known that the loss itself is sufficient for establishing optimal estimation consistency, while the first order information of model estimation is not essential to estimation consistency. See Bickel et al. (2009) for a work on high dimensional linear models and Raskutti et al. (2012) for a study of full non-parametric regression models. To the best of our knowledge, our proof is the first one that provides optimal estimation error for the high dimensional PLM only using the loss itself.

Note that from the proof on our debiased averaging estimation, it is seen that the nonparametric component in PLM significantly affects the estimation error of the averaging parametric estimator. Hence, error bound of our distributed parametric estimator also depends on the functional complexity of nonparametric components. This observation differs from the existing results obtained under the non-distributed setting (Hardle and Liang, 2007; Müller and van de Geer, 2015). Theorem 1 indicates that, under the ultra-high dimensional setting (where the error of parametric estimation dominates the nonparametric one), the parametric estimation with the optimal rate can be guaranteed with an appropriate splitting number $m$, which does not depend on the complex parameter of the nonparametric component. Otherwise, the optimal parametric rate is also guaranteed with a smoothness-dependent $m$.

Our second contribution is to propose a novel debiased distributed estimation for the sparse PLM under the big-data setting, in that simply averaging cannot reduce the estimation bias in contrast to the local estimators. To our knowledge, this is the first work that considers distributed problems on the high dimensional PLM. In fact, our study in this paper is related to the previous work of Zhao et al. (2016), where they considered the naive averaging strategy for the PLM with non-sparse coefficients and fixed dimensions. Hence, their work differs from the current paper in terms of problem setup and methodological strategy. Although the debiasing technology has been employed for the sparse linear regression (Lee et al., 2017), analyzing the debiased distributed estimation for PLM is more challenging, mainly because the nonparametric component affects the error level of the averaging estimation. To handle this problem, we apply some abstract operator theory to provide upper bounds of this approximation error in RKHS. By contrast, some existing related results (Wahba, 1990; Ni et al., 2009) depend on some strong assumptions on data sampling, for example, $T_i$'s are deterministically drawn from $[0, 1]$ such that $\int_0^{T_i} u(t)d(t) = i/n$, where $u(t)$ is a continuous and positive function. From our simulated results, we see the estimation error of the averaging debiased parametric estimator is comparable to that of the centralized M-estimator, while that of the naive averaged parametric estimator is much worse.

The rest of the paper is organized as follows. In Section 2, we provide some background on kernel spaces and propose a debiased averaging parametric estimator based on each local estimate (2). Section 3 is devoted to the statement of our main results and discussion of their consequences. In Section 4, for local parametric estimation, we present general upper bounds on the estimation error. Section 5 contains the technical details, and some useful

lemmas are deferred to the Appendix. Some simulation experiments are reported in Section 6, and we conclude in Section 7.

**Notations**. In the following, for a vector $Z = (z_1, ..., z_p)'$, we use $\|Z\|_1$, $\|Z\|_2$ to represent $\ell_1$ and $\ell_2$-norm in the Euclidean space respectively, and also $\|Z\|_\infty = \max_{j\in[p]} |z_j|$. For a matrix $A$, $\|A\|_2$ denotes the spectral norm. For a function $f$ defined on $\mathcal{Z}$ and a given data $(Z_i)_{i=1}^n$ drawn from the underlying distribution $\rho$ defined on $\mathcal{Z}$, let $\|f\|_2 := \sqrt{\mathbb{E}[f^2(Z)]}$ be the $L_2(\rho)$-norm for any square-integrable function $f$. We use $\|f\|_n$ to denote the empirical $L_2(\rho)$-norm, i.e. $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(Z_i)^2$. With a slight abuse of notation, for a $n$-dimensional vector $v = (v_1, ..., v_n)'$, we still define $\|v\|_n^2 = \frac{1}{n}\sum_{i=1}^n v_i^2$ by convention. For sequences $f(n)$, $g(n)$, $f(n) = \Omega(g(n))$ means that there is some constant $c$, such that $f(n) \geq cg(n)$, and $f(n) = O(g(n))$ means that $f(n) \leq c'g(n)$ for an absolute constant $c'$ with probability approaching one. The symbols $C$, $c$ with various subscripts are used to denote different constants. For $q \in \mathbb{N}^+$, we write $[q] := \{1, ..., q\}$.

## 2. Background and The Proposed Estimator

We begin with some background on RKHSs, and then formulate a profiled Lasso-type approach equivalent to the doubly regularized one (2). Based on a gradient-induced debiasing and an approximation of the inverse weighted covariate matrix, we propose the debiased averaging parametric estimator for PLM.

### 2.1 Reproducing kernel Hilbert space

Given a compact subset $\mathcal{T} \in \mathbb{R}^q$ and a probability measure $\rho_T$, we define a symmetric nonnegative kernel function $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, meaning that $\sum_{i,j=1}^m a_i a_j K(t_i, t_j) \geq 0$ for any $a_i \in \mathbb{R}$ and $t_i \in \mathcal{T}$. A Hilbert space $\mathcal{H}_K$ with the dot product $\langle\cdot,\cdot\rangle_K$ is associated with $K$, and such that (i) for any $t \in \mathcal{T}$, $K_t(\cdot) := K(\cdot, t) \in \mathcal{H}_K$; (ii) the reproducing property holds, e.g. $f(t) = \langle f, K_t\rangle_K$ for all $f \in \mathcal{H}_K$. Actually, under suitable conditions, it is shown that the kernel function $K$ and $\mathcal{H}_K$ are determined by each other (Aronszajn, 1950). Without loss of generality, we assume that $\kappa := \sup_{t\in\mathcal{T}} |K(t, t)| \leq 1$, and such a condition includes the Gaussian kernel and the Laplace kernel as special cases.

The reproducing property of RKHS plays an important role in theoretical analysis and numerical optimization for any kernel-based method. Specially, this property implies that $\|f\|_\infty \leq \kappa\|f\|_K \leq \|f\|_K$ for all $f \in \mathcal{H}_K$. Moreover, by Mercer's theorem, a kernel $K$ defined on a compact subset $\mathcal{T}$ admits the following eigen-decomposition:

$$K(t, t') = \sum_{\ell=1}^{\infty} \mu_\ell \phi_\ell(t)\phi_\ell(t'),\ t, t' \in \mathcal{T},$$

where $\mu_1 \geq \mu_2 \geq \cdots > 0$ are the eigenvalues and $\{\phi_\ell\}_{\ell=1}^\infty$ is an orthonormal basis in $L_2(\rho_T)$. The decay rate of $\mu_\ell$ fully characterizes the complexity of the RKHS induced by the kernel $K$, and has close relationships with various entropy numbers; see Steinwart and Christmann (2008) for details. Based on this, we define the quantity:

$$\mathcal{Q}_n(r) = \frac{1}{\sqrt{n}}\Big[\sum_{\ell=1}^{\infty} \min\{r^2, \mu_\ell\}\Big]^{1/2}, \quad \forall\, r > 0.$$

Let $\nu_n$ be the smallest positive solution to the inequality: $40\nu_n^2 \geq \mathcal{Q}_n(\nu_n)$, where 40 is only a technical constant. Then, due to the high dimensional effect on the nonparametric estimation for PLM, we introduce the following quantity related to the convergence rates of semi-parametric estimate:

$$\gamma_n := \max\Big\{\nu_n, \sqrt{\frac{\log p}{n}}\Big\}.$$

### 2.2 The Debiased Estimator

For the $l$-th machine, define $\mathbb{X}^{(l)} = (X_1^{(l)}, ..., X_n^{(l)})'$, $\boldsymbol{\epsilon}^{(l)} = (\epsilon_1^{(l)}, ..., \epsilon_n^{(l)})'$, $\mathbf{Y}^{(l)} = (Y_1^{(l)}, ..., Y_n^{(l)})'$ and $\mathbf{f}^{(l)} = (f(T_1^{(l)}), ..., f(T_n^{(l)}))'$. $\mathbb{K}^{(l)}$ is a semi-definite $n \times n$ matrix whose entries are $(K(T_i^{(l)}, T_j^{(l)}))_{i,j=1}^n$. The partially linear model (1) can then be written as $\mathbf{Y}^{(l)} = \mathbb{X}^{(l)}\boldsymbol{\beta}^* + (\mathbf{f}^*)^{(l)} + \boldsymbol{\epsilon}^{(l)}$. By the reproducing property of RKHS (Aronszajn, 1950), the nonparametric minimizer of programme (2) has the form $f = \sum_{i\in S_l} a_i K(X_i, \cdot)$ and particularly $\mathbf{f}^{(l)} = \mathbb{K}^{(l)}\mathbf{a}$. Hence we can write $\mathcal{L}^{(l)}(\boldsymbol{\beta}, f)$ as

$$\mathcal{L}^{(l)}(\boldsymbol{\beta}, \mathbf{a}) := \frac{1}{2n}\|\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta} - \mathbb{K}^{(l)}\mathbf{a}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\mathbf{a}^T\mathbb{K}^{(l)}\mathbf{a}. \tag{3}$$

Given $\lambda_1, \lambda_2$ and $\boldsymbol{\beta}$, the first order optimality condition for convex optimization yields the solution

$$\hat{\mathbf{a}}^{(l)} = (n\lambda_2\mathbb{I} + \mathbb{K}^{(l)})^{-1}(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta}), \tag{4}$$

and $\mathbb{A}^{(l)}(\lambda_2) = \mathbb{K}^{(l)}(\lambda_2 n\mathbb{I} + \mathbb{K}^{(l)})^{-1}$ is equivalent to the linear smoother matrix in Heckman (1986). Indeed, $\mathbb{A}^{(l)}(\lambda_2)$ can be replaced by arbitrary smoother for specific purposes. Plugging $\hat{\mathbf{f}}^{(l)} = \mathbb{K}^{(l)}\hat{\mathbf{a}}^{(l)}$ into (2), we can obtain a penalized problem only involving $\boldsymbol{\beta}$:

$$\mathcal{Q}^{(l)}(\boldsymbol{\beta}) := \frac{1}{2n}(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta})'\big(\mathbb{I} - \mathbb{A}^{(l)}(\lambda_2)\big)(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1, \tag{5}$$

and the quadratic term in $\mathcal{Q}^{(l)}(\boldsymbol{\beta})$ is called the profiled least squares in the literature. Note that $\mathbb{I} - \mathbb{A}^{(l)}(\lambda_2) = \big(\mathbb{I} + \mathbb{K}_t^{(l)}/(\lambda_2 n)\big)^{-1}$ is a nonnegative definite smoothing matrix.

Since the gradient vector of the empirical risk $\frac{1}{2n}(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta})'\big(\mathbb{I} - \mathbb{A}^{(l)}(\lambda_2)\big)(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$ is

$$\frac{1}{n}(\mathbb{X}^{(l)})'\big(\mathbb{I} - \mathbb{A}^{(l)}(\lambda_2)\big)(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\hat{\boldsymbol{\beta}}),$$

which is just a sub-gradient of $\lambda_1\|\cdot\|_1$ at $\hat{\boldsymbol{\beta}}$ by the classical KKT conditions. By adding a term proportional to the sub-gradient of the empirical risk for debiasing, any debiased Lasso estimator compensates for the bias incurred by regularization. To be precise, motivated by the idea in Javanmard and Montanari (2014), the debiased estimator from the $l$-th subsample with respect to the Lasso estimator $\hat{\boldsymbol{\beta}}^{(l)}$ is given by

$$\check{\boldsymbol{\beta}}^{(l)} := \hat{\boldsymbol{\beta}}^{(l)} + \frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'\big(\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2)\big)(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\hat{\boldsymbol{\beta}}^{(l)}),$$

where $\hat{\Theta}^{(l)}$ is an approximate inverse to the weight empirical covariance matrix $\tilde{\Sigma}_l := \frac{1}{n}(\tilde{\mathbb{X}}^{(l)})'\tilde{\mathbb{X}}^{(l)}$ on the design matrix $\tilde{\mathbb{X}}^{(l)} := \big(\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2)\big)^{1/2}\mathbb{X}$. Here $\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2) := \big(\mathbb{I} +$

$\mathbb{K}_t^{(l)}/(\lambda_2)\big)^{-1}$ is viewed as an undersmoothed version of $\mathbb{I} - \mathbb{A}^{(l)}(\lambda_2)$ for technical convenience. Note that we drop the dependence on $\lambda_2$ of $\tilde{\mathbb{X}}^{(l)}$ for simplicity. By the debiasing technique, Javanmard and Montanari (2014) proved that the bias of the debaised estimator is of a smaller order than its variance, and thus statistical inference such as asymptotic normality can be tractable.

Note also that the choice of $\hat{\Theta}^{(l)}$ is crucial to the performance of the debiased estimator, and some feasible algorithms for forming $\hat{\Theta}^{(l)}$ have been proposed recently by Javanmard and Montanari (2014) and van der Geer et al. (2014). Thus, the averaged parametric estimator $\bar{\boldsymbol{\beta}}$ by combining the debiased estimators from all the subsamples is given by

$$\bar{\boldsymbol{\beta}} = \frac{1}{m}\sum_{l=1}^{m} \check{\boldsymbol{\beta}}^{(l)}. \tag{6}$$

In this paper, we employ the nodewise Lasso estimator proposed in van der Geer et al. (2014) to form $\hat{\Theta}^{(l)}$. More precisely, for some $j \in [p]$, the $l$-th machine solves

$$\hat{\theta}_j := \arg\min_{\theta \in \mathbb{R}^{p-1}} \frac{1}{2n}\|\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\theta\|_2^2 + \lambda^{(j)}\|\theta\|_1, \tag{7}$$

where $\tilde{\mathbb{X}}_{l,-j} \in \mathbb{R}^{n\times(p-1)}$ is $\tilde{\mathbb{X}}^{(l)}$ without the $j$-th column $\tilde{X}_{l,j}$. Then we can define a non-normalized covariance matrix by

$$\hat{C}_l := \begin{bmatrix} 1 & -\hat{\theta}_{1,2} & \cdots & -\hat{\theta}_{1,p} \\ -\hat{\theta}_{2,1} & 1 & \cdots & -\hat{\theta}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\theta}_{p,1} & -\hat{\theta}_{p,2} & \cdots & 1 \end{bmatrix},$$

where $\hat{\theta}_{j,k}$ is the $k$-th element of $\hat{\theta}_j$, indexed by $k \in \{1, ..., j-1, j+1, ..., p\}$. To scale the rows of $\hat{C}_l$, we define a diagonal matrix $\hat{T}_l := \mathrm{diag}(\hat{\tau_1}, ..., \hat{\tau_p})$ by

$$\hat{\tau_j} = \big(\frac{1}{n}\|\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\|_2^2 + \lambda^{(j)}\|\hat{\theta}_j\|_1\big)^{\frac{1}{2}},$$

and based on this, we form $\hat{\Theta}^{(l)} = \hat{T}_l^{-2}\hat{C}_l$.

In order to show that $\hat{\Theta}^{(l)}$ is an approximate inverse of $\tilde{\Sigma}_l$, the first order optimality conditions of (7) is applied to yield

$$\begin{aligned} \hat{\tau_j}^2 &= \frac{1}{n}\|\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\|_2^2 + \lambda^{(j)}\|\hat{\theta}_j\|_1 \\ &= \frac{1}{n}\|\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\|_2^2 + \frac{1}{n}\big(\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\big)^T\tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j \\ &= \frac{1}{n}\big(\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\big)'\tilde{X}_{l,j}, \;\; j \in [p]. \end{aligned}$$

Let $\hat{\Theta}_{j,\cdot}^{(l)}$ be the $j$-th row of $\hat{\Theta}^{(l)}$. It follows from the above equation and the definition of $\hat{\Theta}^{(l)}$ that

$$\frac{1}{n}\hat{\Theta}_{j,\cdot}^{(l)}(\tilde{\mathbb{X}}^{(l)})'\tilde{X}_{l,j} = \frac{1}{n\hat{\tau_j}^2}\big(\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\big)'\tilde{X}_{l,j} = 1, \;\; j \in [p].$$

Then applying the optimality condition of (7) again yields

$$\frac{1}{n}\big\|\hat{\Theta}_{j,\cdot}^{(l)}(\tilde{\mathbb{X}}^{(l)})'\tilde{\mathbb{X}}_{-j}^{(l)}\big\|_\infty = \frac{1}{\hat{\tau_j}^2}\big\|\frac{1}{n}\big(\tilde{X}_{l,j} - \tilde{\mathbb{X}}_{l,-j}\hat{\theta}_j\big)'\tilde{\mathbb{X}}_{-j}^{(l)}\big\|_\infty \leq \frac{\lambda^{(j)}}{\hat{\tau_j}^2},\ \ j \in [p].$$

Finally, we have

$$\max_{j\in[p]}\big\|\hat{\Theta}_{j,\cdot}^{(l)}\tilde{\Sigma}_l - \mathbf{e}_j\big\|_\infty \leq \max_{j\in[p]}\big\{\lambda^{(j)}/\hat{\tau_j}^2\big\}. \tag{8}$$

We would like to remark that, allowing for moderately high dimensional cases, similar arguments as Lemma 1 in Speckman (1988) with (8) tell us that $\hat{\tau_j}$'s converges to the corresponding eigenvalues of $\Xi$. Here $\mathbf{e} = X - \mathbb{E}[X|T]$ and we write $\Xi := \mathbb{E}[Var(\mathbf{e}|T)]$. Moreover, the positive definiteness of $\Xi$ is a standard assumption for obtaining semi-parametric efficient estimation; see Speckman (1988), Xie and Huang (2009) and Zhao et al. (2016) for more details. In the ultra-high dimensional setting, the underlying sparse structure of $\Xi$ is required to ensure consistency. We impose a related assumption (Assumption E) in Section 3.

We also remark that, if each entry of $\mathbb{E}[X|T]$ belongs to $\mathcal{H}_K$, it is seen that $\mathbb{E}[X|T] = \arg\min_{\mathbf{f}\in\mathcal{H}_K^p}\mathbb{E}[\|X-\mathbf{f}\|^2]$ is the orthogonal projection of $X$ onto $\mathcal{H}_K^p$ and residual error is expressed as $\mathbf{e} := X - \mathbb{E}[X|T]$. Accordingly, the empirical quasi-projection of $X$ onto $\mathcal{H}_K^p$ is defined as $\arg\min_{\mathbf{f}\in\mathcal{H}_K^p}\frac{1}{n}\sum_{i=1}^n\|X_i - \mathbf{f}(T_i)\|^2 + \lambda_2\|\mathbf{f}\|_{\mathcal{H}_K^p}^2$, where $\|\mathbf{f}\|_{\mathcal{H}_K^p}^2 := \sum_{j=1}^p\|f_j\|_K^2$ for any $\mathbf{f} = (f_1, ..., f_p)$. By the reproducing property of Mercer kernel, we observe that the empirical residual error is just the design matrix $\tilde{\mathbb{X}}$. Therefore, it makes sense that $\Xi$ is the population version of $\tilde{\Sigma}$ given that $\lambda_2 \to 0$ and $T_i$'s are fixed. In fact, the detailed proof about $\tilde{\Sigma} \to \Xi$ in partial linear models has been given in Lemma 1 of Appendix A in Speckman (1988).

In addition, by (4), we see that the nonparametric solution has a closed form in terms of the parametric coefficients, and the distributed estimation of the nonparametric components in (2) could also be formulated. In general, the parametric estimation in PLM is more difficult than the nonparametric part, requiring more refined theoretical analysis.

## 3. Theoretical Results

In this section, we first present several assumptions used in our theoretical analysis and introduce further notations. Thereafter, the assumptions are explained explicitly and the main results are stated.

**Assumption A** (Model Specification). For partially linear model (1), we assume that (i) $\boldsymbol{\beta}^*$ is sparse with sparsity $s^*$, and the nonparametric component $f^*$ is a multivariate zero-mean function in the RKHS defined on $\mathcal{T}$; (ii) the noise terms $\epsilon_i$'s are independent normal variables, and also uncorrelated with covariates $(X_i, T_i)$.

**Assumption B** (Covariates Behaviors). (i) The covariate $X$ in $\mathbb{R}^p$ is bounded uniformly, that is, $|X_{\cdot j}| \leq C_0$ for all $j \in [p]$. (ii) The largest eigenvalue of the covariance matrix $\Sigma = \mathbb{E}[XX']$ is finite, denoted by $\Gamma_{\max}$.

Gaussian error assumption is quite strong, but standard in the literature. In general, this condition can be easily relaxed to sub-Gaussian errors. It is worth noting that we allow correlations between $X$ and $T$. The assumption that the target function belongs to the

RKHS is frequently used in machine learning and statistics literature, see Steinwart and Christmann (2008); Raskutti et al. (2012); Zhao et al. (2016) and many others. A bound on the $X$-values is a more restrictive assumption than the sub-Gaussian tails, and we use it for technical reasons.

To estimate the parametric and nonparametric parts respectively, we need some conditions concerning correlations between $X$ and $T$. For each $j \in [p]$, let $\Pi_T^{(j)}$ be the projection of $X^{(j)}$ onto $\mathcal{H}_K$. That is, $\Pi_T^{(j)} = g_j^*(T)$ with

$$g_j^* = \arg\min_{g \in \mathcal{H}_K} \mathbb{E}_{X^{(j)},T}[(X^{(j)} - g(T))^2].$$

We write $\Pi_{X|T} = (\Pi_T^{(1)}, ..., \Pi_T^{(p)})'$ and $X_T = X - \Pi_{X|T}$. Each function $g_j^*$ can be viewed as the best approximation of $\mathbb{E}[X^{(j)}|T]$ within $\mathcal{H}_K$. In the extreme case, when $X$ is uncorrelated with $T$, we get $\Pi_{X|T} = 0$. The following condition ensures that there is enough information in the data to identify the parametric coefficients, which has been imposed in Yu et al. (2011); Müller and van de Geer (2015).

**Assumption C** (Mutual Correlation). (i) The smallest eigenvalue $\Lambda_{\min}$ of $\mathbb{E}[X_T X_T']$ is positive. (ii) The largest eigenvalue of $\Sigma_\pi = \left(\langle g_k^*, g_l^* \rangle_K\right)_{k,l=1,...,p}$ is finite with high probability, denoted by $\Lambda_K$.

**Assumption D** (Spectral and sup-norm assumption). For some $0 < \tau < 1$, there exist two universal constants $C_1, C_2 > 0$, such that
(i) $\mu_\ell \le C_1 \ell^{-1/\tau} and$ (ii) $\|g\|_\infty \le C_2 \|g\|_2^{1-\tau} \|g\|_K^\tau, \quad \forall\, g \in \mathcal{H}_K.$

This assumption is satisfied if the RKHS is a Sobolev space or is continuously embeddable in a Sobolev space. For instance, the RKHSs of Gaussian kernels are continuously embedded in all Sobolev spaces, and thus satisfy sup-norm assumption. More general sufficient conditions to guarantee Assumption D(ii) are related to *real interpolation* shown in Steinwart et al. (2009).

**Assumption E** (Generalized Coherence). Given $\tilde{\Sigma}_l$ defined on the $l$-th subsample, the generalized coherence between $\tilde{\Sigma}_l$ and $\Xi \in R^{p \times p}$ is defined by

$$GC(\tilde{\Sigma}_l, \Xi) = \max_{j \in [p]} \|\tilde{\Sigma}_l \Xi_j' - \mathbf{e}_j\|_\infty.$$

We assume that

$$GC(\tilde{\Sigma}_l, \Xi) = O_p(\sqrt{\log p/n}), \quad \max_{j \in [p]} \left\|\Xi_j^{-1} - \hat{\Theta}_j^{(l)}\right\|_1 = O_p(s_{max}\sqrt{\log p/n}),$$

where $s_{\max}$ is the largest sparsity of the rows of $\Xi^{-1}$.

The preceding definition is viewed as a generalization of the standard coherence appearing in the compressed sensing literature. It is worth noting that, $\tilde{\mathbb{X}}^{(l)}$ defined as above can be viewed as an adjustment of $\mathbb{X}$ for dependence on $T$. Hence, Javanmard and Montanari (2014) and Lemma 2 in Lee et al. (2017) for purely linear models are also applicable to our semi-parametric setting for any given $T$. That is, the GC condition is fulfilled when $\Xi$ is strictly positive and $X - \mathbb{E}[X|T]$ is a sub-Gaussian random vector. Besides, Theorem 2.4 in van der Geer et al. (2014) shows that $\hat{\Theta}_j^{(l)}$ converges to $\Xi_j^{-1}$ at the usual convergence rate of the lasso.

We are now ready to state our main results concerning the debiased averaging estimator (6). It indicates that the averaging estimator achieves the convergence rate of the centralized doubly regularized estimator in some cases, as long as the dataset is not split across too many machines.

**Theorem 1** *Suppose that Assumptions A-E hold. Let* $\lambda^0 := \lambda^{(1)} = ... = \lambda^{(p)}$ *for* $\lambda^{(j)}$ *given in* (7)*. We consider the debiased averaging parametric estimator defined by* (6)*, with suitable parameters constraints:*

$$\lambda^0 \simeq \lambda_1 \simeq \sqrt{\log p/n}, \quad \lambda_2 \simeq n^{-\frac{1}{\tau+1}}.$$

*Then we have*

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty = O_p\Big((s^* + s_{max})\frac{\log p}{n} + \sqrt{\lambda_{\max}(\hat{\Theta}^{(l)})}n^{-\frac{(2+\tau)}{2(1+\tau)}} + \sqrt{\max_j\{(\hat{\Theta}^{(l)})_{jj}\}}\sqrt{\frac{\log p}{N}}\Big).$$

In the following, we assume that $s^* \simeq s_{max}$ for notational simplicity. The rate of Theorem 1 may be interpreted as the sum of estimation error of the parametric part $O\big(s^*\frac{\log p}{n}\big)$, the influence of the nonparametric component in PLM $O\big(n^{-\frac{(2+\tau)}{2(1+\tau)}}\big)$, and the total noise error $O\big(\sqrt{\frac{\log p}{N}}\big)$. When $p$ grows exponentially with $n$ (e.g. $\log p = O(n^r)$ with $\frac{\tau}{2(1+\tau)} < r < 1$), $\big(n^{-\frac{(2+\tau)}{2(1+\tau)}}\big)$ is negligible compared to $\big(s^*\frac{\log p}{n}\big)$ and the error of $\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty$ is of the order $s^*\big(\sqrt{\frac{\log p}{N}}\big)$ by choosing $m \leq \sqrt{N/\log p}$. Otherwise, if the term $\big(n^{-\frac{(2+\tau)}{2(1+\tau)}}\big)$ dominates the parametric error, the total error $s^*\big(\sqrt{\frac{\log p}{N}}\big)$ can be also achieved by choosing $m \leq N^{\frac{1}{2+\tau}}/((\log p)^{\frac{1+\tau}{2+\tau}})$. Up to the logarithmic term, the smallest number of data partition (corresponding to the worst case $\tau = 1$) is $m = N^{\frac{1}{3}}$. It is interesting to note that the number of splits may be larger as the nonparametric function becomes smoother. In summary, the partition-based parametric estimator achieves the statistical minimax rate over all estimators using the set of $N$ samples.

The result in Theorem 1 also indicates that the choice of the number $m$ of subsampled datasets does not rely on $s^*$ , which means that $m$ is adaptive to the sparsity parameter.

As an illustration, for the RKHS with eigenvalues which decay at a rate $\mu_\ell = (1/\ell)^{2\alpha}$ for some parameter $\alpha > 1/2$, this type of scaling covers the case of Sobolev spaces and Besov spaces. In this case we can check that $\nu_n = n^{-\frac{\alpha}{2\alpha+1}}$.

**Corollary 2** *Under the same conditions as in Theorem 1 within a Sobolev space with* $\alpha$ *derivatives. When* $p$ *diverges polynomially fast with the local sample size* $n$*, by choosing* $\lambda^0 \simeq \lambda_1 \simeq \sqrt{\log p/n}$ *and* $\lambda_2 \simeq n^{-\frac{2\alpha}{2\alpha+1}}$*, the estimation error of the parametric estimator is bounded by*

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty = O_p\big(s^*\sqrt{\frac{\log p}{N}}\big),$$

*provided that the number of local machines* $m$ *satisfy the bound* $m \leq N^{\frac{2\alpha}{4\alpha+1}}$.

Our above arguments imply the relation between the split number $m$ and the smoothness parameter $\alpha$. We remark that, the main challenge in deriving the optimal minimax rates for our averaging parametric estimator comes from the negative influence of the nonparametric component in PLM, and this differs from the semi-parametric literature under the non-distributed setting; see, for example, Chen (1988) and Zhu (2017).

## 4. Estimation on Local Estimators

This section provides general upper bounds on $\|X_T'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\|_2$ and $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1$ for the standard PLM (1). We now state the sketch of our main proof. First, we provide a crucial inequality that characterizes the relation between the parametric estimator and the nonparametric estimator (see Theorem 2 below). Second, under the same conditions as in Theorem 2, Proposition 1 shows that our estimators are bounded uniformly with high probability. Thus, our final results follow immediately. The proof idea we adopt is to avoid the use of the first-order information for our programme (2). In this paper, we are particularly interested in estimating $\boldsymbol{\beta}^*$ when it is sparse with diverging dimensions.

In what follows, we write $\Delta_\beta = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and $\Delta_f = f - f^*$, and particularly $\widehat{\Delta}_\beta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $\widehat{\Delta}_f = \hat{f} - f^*$.

**Theorem 3** *Suppose that Assumptions A-D hold, and we consider the doubly regularized estimator $(\hat{\boldsymbol{\beta}}, \hat{f})$ defined by (2), with the following parameters constraints*

$$\lambda_1/2 \geq r_n + 2\big(2N_0\sqrt{(C_0+\Pi_{max}+2)/n} + (C_0+\Pi_{max})\sqrt{r_n/n}\big)\|\widehat{\Delta}_f\|_K,\ \ \lambda_2\Gamma_{\max}/\Lambda_{\min} \leq 1/2,$$

*then for any $r_n > 0$ we have*

$$1/2\|X_T'\widehat{\Delta}_\beta\|_2^2 + \lambda_1\|\widehat{\Delta}_\beta\|_1 \leq 3\|\widehat{\Delta}_\beta\|_1^2 r_n + \big(4\sqrt{s^*/\Lambda_{\min}}\lambda_1 + 2\lambda_2\|f^*\|_K\sqrt{\Gamma_{\max}/\Lambda_{\min}}\big)\|X_T'\widehat{\Delta}_\beta\|_2 + 2\lambda_2\sqrt{\Lambda_K/\Lambda_{\min}}\|X_T'\widehat{\Delta}_\beta\|_2\|\widehat{\Delta}_f\|_K, \tag{9}$$

*with probability at least*

$$1 - e\cdot p\exp\big(-\frac{cnr_n^2}{\max_{j\in[p]}C_{K,j}^2}\big) - 2e\cdot p^2\exp\Big(-\frac{cnr_n^2}{4(C_0+\Pi_{max})^4}\Big) - 2p\exp(-nr_n^2).$$

*Here $\Pi_{max} = \max_{j\in[p]}\|g_j^*\|_\infty$. $C_{j,K}$ and $N_0$ are two absolute constants appearing in the proof of Theorem 2 and Lemma 7 respectively.*

The proof of Theorem 3, contained in Section 5, constitutes one main technical contribution of this paper. From the result of Theorem 3, it is seen that the quadratic term $\|\widehat{\Delta}_\beta\|_1^2$ appears in the right hand side of (9). Hence, this result is useful only if upper bounds of $\|\widehat{\Delta}_\beta\|_1$ and $\|\widehat{\Delta}_f\|_K$ can be proved to be bounded uniformly in advance. Proposition 1 attempts at solving such a problem. Recall that a standard technical proof for analyzing (9) is to first construct some event, and then prove that the desired results hold under the event, as well as that the event occurs with high probability (see Müller and van de Geer (2015) for details). However, their results are based on the use of the first order optimization. We also notice that the restricted strong convexity (Raskutti et al., 2012) and restricted eigenvalues

constants (Bickel et al., 2009) are two key tools to derive sharp error bounds of the oracle results. These mentioned techniques can be used for the two-step estimation for PLM (Zhu, 2017). Nevertheless, it seems quite difficult to apply for our one-step approach.

**Proposition 4** *Suppose that Assumptions A-D hold, with* $\lambda_1 \simeq \sqrt{\log(2p)/n}$ *and* $\lambda_2 \simeq n^{-\frac{1}{\tau+1}}$. *Then, we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \sqrt{s^*} O_p\Big(\frac{n^{\frac{1}{2}-\frac{1}{1+\tau}}}{\sqrt{\log p}}\Big),$$

*and*

$$\|\hat{f} - f^*\|_K = O_p(1).$$

The results in Proposition 4 follow easily from Lemma 6 below with suitable choices of parameters. Here we omit the proof details for Proposition 1. Moreover, Lemma 13 and Lemma 14 in the Appendix guarantee that the event in Lemma 6 holds with probability tending to 1.

Note that $0 < \tau < 1$ implies $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p(1)$ provided that $s^*$ is rather small as compared to $n$ and $p$. Thus, combining the results of Theorem 2 and Proposition 1, the following theorem follows immediately from Theorem 3.

**Theorem 5** *Suppose the same conditions of Theorem 3 hold with* $s^* = O(n^{\frac{1}{1+\tau}-\frac{1}{2}})$. *By choosing* $\lambda_1 \simeq \sqrt{\log p/n}$ *and* $\lambda_2 \simeq n^{-\frac{1}{\tau+1}}$, *we have*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O\Big(\sqrt{\frac{s^* \log p}{n}}\Big)$$

*and*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O\Big(s^* \sqrt{\frac{\log p}{n}}\Big)$$

*with probability at least* $1 - c_1 p^{-1}$ *for some universal constant* $c_1 > 0$.

Remark that, our theoretical results regarding the properties of the parameter estimators in (2) are non-asymptotic. Indeed, estimation error of the nonparametric estimator is also obtained from Lemma 1 below and the result of Theorem 5, but this is not our current focus and we omit the details.

We notice that these two rates are of the same order as the standard Lasso for linear models (see, for example, Bickel et al. (2009) and van der Geer et al. (2014)). In other words, despite the presence of the non-parametric part, the parametric part can be estimated with parametric rate under regularity conditions.

## 5. Proofs

In this section, we provide the detailed proofs of Theorems 1-2. First of all, we give the proof of each local parametric estimate, which is one of the key ingredients for obtaining the oracle rates of the averaging estimator based on the entire data.

### 5.1 Quantitative Relation Between Local Estimators

In this subsection, we focus on theoretical analysis on each local machine $l$ ($l \in [m]$) in (2). For all the symbols and numbers, we drop their dependence on $l$ for notational simplicity.

**Proof of Theorem 3**. By the definition of $(\hat{\boldsymbol{\beta}}, \hat{f})$ in (2), we have $\mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{f}) \leq \mathcal{L}(\boldsymbol{\beta}^*, \hat{f} + \Pi'_{X|T}\widehat{\Delta}_\beta)$, which means that

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{Y} - X'\hat{\boldsymbol{\beta}} - \hat{f}\|_n^2 + \lambda_1\|\hat{\boldsymbol{\beta}}\|_1 + \frac{\lambda_2}{2}\|\hat{f}\|_K^2 \\
&\leq \frac{1}{2}\|\mathbf{Y} - X'\boldsymbol{\beta}^* - (\hat{f} + \Pi'_{X|T}\widehat{\Delta}_\beta)\|_n^2 + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \frac{\lambda_2}{2}\|\hat{f} + \Pi'_{X|T}\widehat{\Delta}_\beta\|_K^2.
\end{aligned}$$

Since $\|\hat{\boldsymbol{\beta}}\|_1 = \|\hat{\boldsymbol{\beta}}_S\|_1 + \|\hat{\boldsymbol{\beta}}_{S^c}\|_1$ and $\|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}^*_S\|_1$ by sparsity assumption in model (1), by the triangle inequality, the last inequality implies that

$$\begin{aligned}
&\|X'_T\widehat{\Delta}_\beta\|_n^2 + 2\lambda_1\|\widehat{\Delta}_\beta\|_1 \\
&\leq 2\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_T)'_i\widehat{\Delta}_\beta\Big| + 2\Big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_f(T_i)(X_T)'_i\widehat{\Delta}_\beta\Big| + 2\Big|\frac{1}{n}\sum_{i=1}^n (\Pi_{X|T})'_i(\widehat{\Delta}_\beta)(X_T)'_i\widehat{\Delta}_\beta\Big| \\
&+ 4\lambda_1\|(\widehat{\Delta}_\beta)_S\|_1 + \lambda_2(\|\hat{f} + \Pi'_{X|T}\widehat{\Delta}_\beta\|_K^2 - \|\hat{f}\|_K^2) \\
&:= 2\Phi_1 + 2\Phi_2 + 2\Phi_3 + 4\Phi_4 + \Phi_5. \qquad (10)
\end{aligned}$$

We will give tight upper bounds for all the terms $\Phi_i$'s. First, since $(X_i, T_i)$ are independent of $\epsilon_i$ and they are bounded and Gaussian respectively, $\mathbb{E}[\epsilon_i(X_T)_{ij}] = 0$ and $\epsilon_i(X_T)_{ij}$ is sub-Gaussian, whose sub-Gaussian norms are upper bounded, denoted by $C_{K,j}$, which depends on the Gamma function and $C_0$ and $\|g_j^*\|_\infty$. By the Hoffding-type inequality in Lemma 10 and the union bounds, we have

$$\Phi_1 = \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_T)'_i\widehat{\Delta}_\beta\Big| \leq \max_j \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i(X_T)_{ij}\Big|\|\widehat{\Delta}_\beta\|_1 \leq r_n\|\widehat{\Delta}_\beta\|_1, \qquad (11)$$

with probability at least $1 - e \cdot p\exp\big(-\frac{cnr_n^2}{\max_{j\in[p]} C_{K,j}^2}\big)$.

Next, consider the second term $\Phi_2 = \Big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_f(T_i)(X_T)'_i(\widehat{\Delta}_\beta)\Big|$. Note that

$$\Phi_2 \leq \max_j \Big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_f(T_i)(X_T)_{ij}\Big|\|\widehat{\Delta}_\beta\|_1,$$

and $\mathbb{E}[\widehat{\Delta}_f X_T^{(j)}] = 0$ for each $j \in [p]$ by the definition of projection, and thus the Talagrand's concentration inequality in Hilbert spaces in Lemma 7 below can be used directly. We shall claim that, with probability at least $1 - 2p\exp(-nr_n^2)$, we get

$$\Phi_2 \leq 2\Big(2N_0\sqrt{\frac{C_0 + \Pi_{max} + 2}{n}} + (C_0 + \Pi_{max})\sqrt{\frac{r_n}{n}}\Big)\|\widehat{\Delta}_f\|_K\|\widehat{\Delta}_\beta\|_1, \qquad (12)$$

where $N_0$ appearing in Lemma 7 is some absolute constant, independent of $n$ or $p$. In fact, it is trivial if $\widehat{\Delta}_f$ is zero, and thus it suffices to consider non-zero cases. To this end, we

define the function set

$$\mathcal{G} = \Big\{ g(X,T) = \frac{X_T^{(j)} f(T)}{\|f\|_K}, f \in \mathcal{H}_K - \{0\} \Big\},\ j \in [p]$$

and let $\mathbf{Z} = \sup_{g\in\mathcal{G}} \big|\frac{1}{n}\sum_{i=1}^n g(X_i,T_i)\big|$. Here we often drop the dependence on $j$ for simplicity. By the definition of projection, $\mathbb{E}[f(T)X_T^{(j)}] = 0$ for any $f \in \mathcal{H}_K$. Then (38) in Appendix A can be applied to yield

$$\mathbb{E}[\mathbf{Z}] = \mathbb{E}\Big[\sup_{g\in\mathcal{G}} \big|\frac{1}{n}\sum_{i=1}^n g(X_i,T_i)\big|\Big] \le 2\mathrm{R}_n(\mathcal{G}_1), \tag{13}$$

where $R_n(\cdot)$ refers to the Rademacher complexity defined in the Appendix and $\mathcal{G}_1 = \big\{g(X,T) = X_T^{(j)} f(T),\ \|f\|_K = 1\big\}$. Note that, $\|X_T^{(j)}\|_\infty \le C_0 + \|g_j^*\|_\infty$ for any $j$. By the contraction inequality and sub-additivity of Rademacher complexity, we easily obtain

$$\mathrm{R}_n(\mathcal{G}_1) \le \mathrm{R}_n(X_T^{(j)}) + \mathrm{R}_n(\mathbb{B}_K(1)) \le 2\sqrt{\frac{C_0 + \|g_j^*\|_\infty + 2}{n}},$$

where we used the conclusion in Lemma 9. It immediately follows from (13) that

$$\mathbb{E}[\mathbf{Z}] \le 4\sqrt{\frac{C_0 + \|g_j^*\|_\infty + 2}{n}}. \tag{14}$$

In addition, by the definition of $\mathcal{G}$, we can take $U = C_0 + \|g_j^*\|_\infty$ and $B = (C_0 + \|g_j^*\|_\infty)^2$ in Lemma 7. As a consequence, following (14) and the derived quantities of $U, B$, we conclude from Lemma 7 and the union bounds that

$$\mathbf{Z} \le 4N_0\sqrt{\frac{C_0 + \Pi_{max} + 2}{n}} + 2(C_0 + \Pi_{max})\sqrt{\frac{r_n}{n}}, \tag{15}$$

with probability at least $1 - p\exp(-nr_n^2)$ for any $r_n \le 1$. Let $f = \widehat{\Delta}_f$, then we complete the proof of (12).

We now turn to the term $\Phi_3$. Following the definition of projection, $\mathbb{E}[\Pi_T^{(k)} X_T^{(j)}] = 0$ for all $j, k \in [p]$, then we have the following decomposition:

$$\Phi_3 = \Big|\frac{1}{n}\sum_{i=1}^n (\Pi_{X|T})_i' \widehat{\Delta}_\beta (X_T)_i' \widehat{\Delta}_\beta\Big| = \Big|\widehat{\Delta}_\beta' \Big(\frac{1}{n}\sum_{i=1}^n (\Pi_{X|T})_i (X_T)_i' - \mathbb{E}[\Pi_{X|T} X_T']\Big)\widehat{\Delta}_\beta\Big|. \tag{16}$$

By simple algebra, we have

$$\Phi_3 \le \|\widehat{\Delta}_\beta\|_1^2 \max_{k,j\in[p]} \Big|\frac{1}{n}\sum_{i=1}^n \big[(\Pi_{X|T})_i (X_T)_i' - \mathbb{E}[\Pi_{X|T} X_T']\big]_{k,j}\Big|,$$

where $(\Pi_T^{(j)} X_T^{(k)})_i - \mathbb{E}[(\Pi_T^{(j)})(X_T^{(k)})]$ are upper bounded by $2\|g_j^*\|_\infty(C_0 + \|g_j^*\|_\infty)$ by assumption. Applying the Hoffding-type inequality in Lemma 10 and using the union bound over $j, k \in [p]$ yield

$$\max_{k,j\in[p]} \Big|\frac{1}{n}\sum_{i=1}^n \big[(\Pi_{X|T})_i (X_T)_i' - \mathbb{E}[\Pi_{X|T} X_T']\big]_{k,j}\Big| \le r_n,$$

with probability at least $1 - e \cdot p^2 \exp\Big( - \frac{cnr_n^2}{4\Pi_{max}^2 (C_0 + \Pi_{max})^2} \Big)$. Thus, we have

$$\Phi_3 \leq \|\widehat{\Delta}_\beta\|_1^2 r_n, \tag{17}$$

with the same probability as above. On the other hand, some simple algebra shows that

$$\Phi_4 = \lambda_1 \|(\widehat{\Delta}_\beta)_S\|_1 \leq \sqrt{s^*}\lambda_1 \|\widehat{\Delta}_\beta\|_2 \leq \sqrt{s^*/\Lambda_{\min}}\lambda_1 \|X_T' \widehat{\Delta}_\beta\|_2, \tag{18}$$

where the last inequality follows from Assumption C.

It remains to consider the term $\Omega_5$. A direct computation yields that

$$\begin{aligned}
\Phi_5 &= \lambda_2 (\|\hat{f} + \Pi_{X|T}' \widehat{\Delta}_\beta\|_K^2 - \|\hat{f}\|_K^2) \\
&\leq \lambda_2 \big( \|\Pi_{X|T}' \widehat{\Delta}_\beta\|_K^2 + 2\|\Pi_{X|T}' \widehat{\Delta}_\beta\|_K \|\widehat{\Delta}_f\|_K + 2\|f^*\|_K \|\Pi_{X|T}' \widehat{\Delta}_\beta\|_K \big) \\
&\leq \lambda_2 \Lambda_K/\Lambda_{\min} \|X_T' \widehat{\Delta}_\beta\|_2^2 + 2\lambda_2 \sqrt{\Lambda_K/\Lambda_{\min}} \|X_T' \widehat{\Delta}_\beta\|_2 \|\widehat{\Delta}_f\|_K \\
&\quad + 2\lambda_2 \|f^*\|_K \sqrt{\Lambda_K/\Lambda_{\min}} \|X_T' \widehat{\Delta}_\beta\|_2
\end{aligned} \tag{19}$$

where the second inequality follows from Assumption B.

Therefore, substituting (11), (12), (17), (18) and (19) into (10), we obtain

$$\begin{aligned}
&\|X_T' \widehat{\Delta}_\beta\|_n^2 + 2\lambda_1 \|\widehat{\Delta}_\beta\|_1 \\
&\leq 2\Big( r_n + 2\big( 2N_0 \sqrt{(C_0 + \Pi_{max} + 2)/n} + (C_0 + \Pi_{max}) \sqrt{r_n/n} \big) \|\widehat{\Delta}_f\|_K \Big) \|\widehat{\Delta}_\beta\|_1 \\
&+ 2\|\widehat{\Delta}_\beta\|_1^2 r_n + \big( 4\sqrt{s^*/\Lambda_{\min}}\lambda_1 + 2\lambda_2 \|f^*\|_K \sqrt{\Lambda_K/\Lambda_{\min}} \big) \|X_T' \widehat{\Delta}_\beta\|_2 \\
&+ \lambda_2 \Lambda_K/\Lambda_{\min} \|X_T' \widehat{\Delta}_\beta\|_2^2 + 2\lambda_2 \sqrt{\Lambda_K/\Lambda_{\min}} \|X_T' \widehat{\Delta}_\beta\|_2 \|\widehat{\Delta}_f\|_K
\end{aligned} \tag{20}$$

with probability at least

$$1 - e \cdot p \exp\big( - \frac{cnr_n^2}{\max_{j \in [p]} C_{K,j}^2} \big) - e \cdot p^2 \exp\Big( - \frac{cnr_n^2}{4\Pi_{max}^2 (C_0 + \Pi_{max})^2} \Big) - 2p \exp(-nr_n^2).$$

We now establish an equivalent relationship between $\|X_T' \widehat{\Delta}_\beta\|_n^2$ and $\|X_T' \widehat{\Delta}_\beta\|_2^2$ in high dimensions. To this end, a direct computation leads to

$$\big| \|X_T' \widehat{\Delta}_\beta\|_n^2 - \|X_T' \widehat{\Delta}_\beta\|_2^2 \big| \leq \|\widehat{\Delta}_\beta\|_1^2 \max_{k,j \in [p]} \Big| \frac{1}{n} \sum_{i=1}^n \big[ (X_T)_i (X_T)_i' - \mathbb{E}[X_T X_T'] \big]_{k,j} \Big|,$$

where $(X_T)_{ij}(X_T)_{ik} - \mathbb{E}[(X_T)_{ij}(X_T)_{ik}]$ are upper bounded by $2(C_0 + \|g_j^*\|_\infty)^2$ by Assumption B. Applying the Hoffding-type inequality in Lemma 10 yields

$$\Big| \frac{1}{n} \sum_{i=1}^n (X_T)_{ij}(X_T)_{ik} - \mathbb{E}[(X_T)_{ij}(X_T)_{ik}] \Big| \leq r_n, \quad \forall\ j, k \in [p],$$

with probability at least $1 - e \cdot \exp\Big( - \frac{cnr_n^2}{4(C_0 + \|g_j^*\|_\infty)^4} \Big)$. Then, the union bounds implies that

$$\big| \|X_T' \widehat{\Delta}_\beta\|_n^2 - \|X_T' \widehat{\Delta}_\beta\|_2^2 \big| \leq r_n \|\widehat{\Delta}_\beta\|_1^2, \tag{21}$$

with probability at least $1-e\cdot p^2\exp\Big(-\frac{cnr_n^2}{4(C_0+\Pi_{max})^4}\Big)$. Then, plugging (21) into (20), we obtain that

$$\begin{aligned}
&\|X_T'\widehat{\Delta}_\beta\|_2^2+2\lambda_1\|\widehat{\Delta}_\beta\|_1\\
&\le 2\Big(r_n+2\big(2N_0\sqrt{(C_0+\Pi_{max}+2)/n}+(C_0+\Pi_{max})\sqrt{r_n/n}\big)\|\widehat{\Delta}_f\|_K\Big)\|\widehat{\Delta}_\beta\|_1\\
&+3\|\widehat{\Delta}_\beta\|_1^2 r_n+\big(4\sqrt{s^*/\Lambda_{\min}}\lambda_1+2\lambda_2\|f^*\|_K\sqrt{\Lambda_K/\Lambda_{\min}}\big)\|X_T'\widehat{\Delta}_\beta\|_2\\
&+\lambda_2\Lambda_K/\Lambda_{\min}\|X_T'\widehat{\Delta}_\beta\|_2^2+2\lambda_2\sqrt{\Lambda_K/\Lambda_{\min}}\|X_T'\widehat{\Delta}_\beta\|_2\|\widehat{\Delta}_f\|_K
\end{aligned}\tag{22}$$

with probability at least

$$1-e\cdot p\exp\big(-\frac{cnr_n^2}{\max_{j\in[p]}C_{K,j}^2}\big)-2e\cdot p^2\exp\Big(-\frac{cnr_n^2}{4(C_0+\Pi_{max})^4}\Big)-2p\exp(-nr_n^2).$$

Based on this and with the choice of

$$\lambda_1/2\ge r_n+2\big(2N_0\sqrt{(C_0+\Pi_{max}+2)/n}+(C_0+\Pi_{max})\sqrt{r_n/n}\big)\|\widehat{\Delta}_f\|_K,\ \lambda_2\Lambda_K/\Lambda_{\min}\le 1/2,\tag{23}$$

then with the same probability as above, we have

$$\begin{aligned}
1/2\|X_T'\widehat{\Delta}_\beta\|_2^2+\lambda_1\|\widehat{\Delta}_\beta\|_1\le 3\|\widehat{\Delta}_\beta\|_1^2 r_n+\big(4\sqrt{s^*/\Lambda_{\min}}\lambda_1+2\lambda_2\|f^*\|_K\sqrt{\Lambda_K/\Lambda_{\min}}\big)\|X_T'\widehat{\Delta}_\beta\|_2\\
+2\lambda_2\sqrt{\Lambda_K/\Lambda_{\min}}\|X_T'\widehat{\Delta}_\beta\|_2\|\widehat{\Delta}_f\|_K.
\end{aligned}\tag{24}$$

This completes the proof of Theorem 2. □

Remark that, given that $X$ has a weak correlation with $T$ (e.g. Assumption C(i) holds), the convergence of the parametric estimator $\hat{\boldsymbol{\beta}}$ is guaranteed with suitable choices of $(\lambda_1,\lambda_2)$, as long as both $\|\widehat{\Delta}_f\|_K$ and $\|\widehat{\Delta}_\beta\|_1$ are bounded uniformly. In other words, the convergence of the non-parametric estimator $\hat{f}$ has no effect on the convergence of the parametric estimator.

Technically, an obvious difference from the existing related proof lies in the start point of the proof. In our analysis, we use $\mathcal{L}(\hat{\boldsymbol{\beta}},\hat{f})\le\mathcal{L}(\boldsymbol{\beta}^*,\hat{f}+\Pi_{X|T}'\widehat{\Delta}_\beta)$ instead of the classical zero-optimization $\mathcal{L}(\hat{\boldsymbol{\beta}},\hat{f})\le\mathcal{L}(\boldsymbol{\beta}^*,f^*)$ (see Müller and van de Geer (2015)). Note also that, a lower bound of $\lambda_2$ is required to ensure that $\|\widehat{\Delta}_f\|_K$ can be bounded.

### 5.2 Boundedness for Local Estimators

This subsection is devoted to establishing rough convergence rates for local estimators in (2). Obviously, this means that both $\|\widehat{\Delta}_f\|_K$ and $\|\widehat{\Delta}_\beta\|_1$ are bounded uniformly under suitable choices of $(\lambda_1,\lambda_2)$. The corresponding proof borrows the techniques from Müller and van de Geer (2015). Let $R>0$ and write $\lambda=(\lambda_1,\lambda_2)$. We define

$$\Omega_{R,\lambda}(\boldsymbol{\beta},f):=\lambda_1\|\boldsymbol{\beta}\|_1/(R\sqrt{\delta_0/2})+\sqrt{\|X'\boldsymbol{\beta}+f(T)\|_2^2+\lambda_2\|f\|_K^2},\quad \boldsymbol{\beta}\in\mathbb{R}^p,\ f\in\mathcal{H}_K,$$

where $\delta_0$ is a fixed small constant, specified in the proof of Lemma 6 below. Define a constrained function set

$$\mathcal{G}(R):=\{g(Z)=X'\boldsymbol{\beta}+f(T):\ \Omega_{R,\lambda}(\boldsymbol{\beta},f)\le R\}.$$

We now introduce two events related to empirical processes theory

$$\mathcal{T}_1(\delta_0, R) := \Big\{(X,T),\ \sup_{g\in\mathcal{G}(R)} \big|\|g\|_n^2 - \|g\|_2^2\big| \le \delta_0 R^2\Big\}$$

and

$$\mathcal{T}_2(\delta_0, R) := \Big\{(X,T,\epsilon),\ \sup_{g\in\mathcal{G}(R)} \big|\boldsymbol{\epsilon}' g/n\big| \le \delta_0 R^2\Big\}.$$

With these notations, define the event

$$\mathcal{T}(\delta_0, R) := \mathcal{T}_1(\delta_0, R) \cap \mathcal{T}_2(\delta_0, R).$$

**Lemma 6** *Suppose that Assumptions A and C hold. If the regularization parameters* $(\lambda_1, \lambda_2)$ *satisfy*

$$2\lambda_1^2 s^*/\Lambda_{\min}^2 \le \delta_0 R^2, \quad 2\lambda_2 \|f^*\|_K^2 \le \delta_0 R^2,$$

*then conditioned on* $\mathcal{T}(\delta_0, R)$*, we have* $\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f) \le R$*.*

**Proof** Let $s = R/(R + \Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f))$, and we define two intermediate components by

$$\tilde{\boldsymbol{\beta}} = s\hat{\boldsymbol{\beta}} + (1-s)\boldsymbol{\beta}^*, \quad \tilde{f} = s\hat{f} + (1-s)f^*.$$

Then we have $\tilde{g}(Z) := X'\tilde{\boldsymbol{\beta}} + \tilde{f}(T) = s\hat{g}(Z) + (1-s)g^*(Z)$, where $\hat{g}(Z) = X'\hat{\boldsymbol{\beta}} + \hat{f}(T)$. Note that

$$\Omega_{R,\lambda}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \tilde{f} - f^*) = s\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f) = \frac{R\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f)}{R + \Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f)} \le R,$$

which means that $\tilde{g} - g^* \in \mathcal{G}(R)$. Based on the above formulation, it suffices to show that $\Omega_{R,\lambda}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \tilde{f} - f^*) \le R/2$.

We now state the details of this claim. By convexity of our objective function (2) and definition of $(\hat{\boldsymbol{\beta}}, \hat{f})$, we have

$$\begin{aligned}
&\frac{1}{2}\|Y - \tilde{g}\|_n^2 + \lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \frac{\lambda_2}{2}\|\tilde{f}\|_K^2 \le \\
&s\big(\frac{1}{2}\|Y - \hat{g}\|_n^2 + \lambda_1\|\widehat{\boldsymbol{\beta}}\|_1 + \frac{\lambda_2}{2}\|\hat{f}\|_K^2\big) + (1-s)\big(\frac{1}{2}\|Y - g^*\|_n^2 + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \frac{\lambda_2}{2}\|f^*\|_K^2\big) \\
&\le \frac{1}{2}\|Y - g^*\|_n^2 + \lambda_1\|\boldsymbol{\beta}^*\|_1 + \frac{\lambda_2}{2}\|f^*\|_K^2.
\end{aligned}$$

Using that $Y = g^* + \boldsymbol{\epsilon}$ in our model (1), the above inequality can be rewritten as

$$\|\tilde{g} - g^*\|_n^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_2\|\tilde{f}\|_K^2 \le 2\boldsymbol{\epsilon}'(\tilde{g} - g^*)/n + 2\lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|f^*\|_K^2.$$

As shown above, $\tilde{g} - g^* \in \mathcal{G}(R)$. Hence, on $\mathcal{T}(\delta_0, R)$, we have

$$\|\tilde{g} - g^*\|_2^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}}\|_1 + \lambda_2\|\tilde{f}\|_K^2 \le 3\delta_0 R^2 + 2\lambda_1\|\boldsymbol{\beta}^*\|_1 + \lambda_2\|f^*\|_K^2.$$

Since $\|\tilde{\boldsymbol{\beta}}\|_1 = \|\tilde{\boldsymbol{\beta}}_S\|_1 + \|\tilde{\boldsymbol{\beta}}_{S^c}\|_1$ and $\|\boldsymbol{\beta}^*\|_1 = \|\boldsymbol{\beta}^*_S\|_1$ by sparsity assumption in model (1), the triangle inequality is applied to imply that

$$\begin{aligned}&\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \|\Pi'_{X|T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\tilde{f} - f^*)\|_2^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}}_{S^c}\|_1 + \lambda_2\|\tilde{f}\|_K^2 \\ &\le 3\delta_0 R^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 + \lambda_2\|f^*\|_K^2. \end{aligned} \tag{25}$$

Using $uv \le u^2 + v^2/4$ for any $u, v \in \mathbb{R}$, we get

$$\begin{aligned}\lambda_1\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 &\le \lambda_1\sqrt{s^*}\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_2 \le \lambda_1\sqrt{s^*}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \frac{\lambda_1\sqrt{s^*}}{\Lambda_{\min}}\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \\ &\le \frac{\lambda_1^2 s^*}{\Lambda_{\min}^2} + \frac{\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{4}. \end{aligned} \tag{26}$$

Meanwhile, observing that $\|\tilde{f}\|_K^2 \ge \frac{1}{2}\|\tilde{f} - f^*\|_K^2 - \|f^*\|_K^2$, we conclude from (25) that

$$\begin{aligned}&\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \|\Pi'_{X|T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\tilde{f} - f^*)\|_2^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}}_{S^c}\|_1 + \lambda_2/2\|\tilde{f} - f^*\|_K^2 \\ &\le 3\delta_0 R^2 + \frac{2\lambda_1^2 s^*}{\Lambda_{\min}^2} + \frac{\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{2} + 2\lambda_2\|f^*\|_K^2. \end{aligned}$$

By our assumptions

$$\frac{2\lambda_1^2 s^*}{\Lambda_{\min}^2} \le \delta_0 R^2, \quad 2\lambda_2\|f^*\|_K^2 \le \delta_0 R^2,$$

we obtain

$$\begin{aligned}&1/2\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \|\Pi'_{X|T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\tilde{f} - f^*)\|_2^2 + \lambda_2/2\|\tilde{f} - f^*\|_K^2 \\ &\le 5\delta_0 R^2. \end{aligned} \tag{27}$$

Hence, by orthogonal decomposition of $\Pi_{X|T}$, this leads to

$$\|\tilde{g} - g^*\|_2^2 + \lambda_2\|\tilde{f} - f^*\|_K^2 \le 10\delta_0 R^2,$$

that is,

$$\left(\|X'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\tilde{f} - f^*)\|_2^2 + \lambda_2\|\tilde{f} - f^*\|_K^2\right)^{1/2} \le (10\delta_0)^{1/2} R. \tag{28}$$

On the other hand, note also that $\|\hat{\boldsymbol{\beta}}_{S^c}\|_1 = \|\hat{\boldsymbol{\beta}}_{S^c} - \boldsymbol{\beta}^*_{S^c}\|_1$ using sparsity assumption again. Then adding $2\lambda_1\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1$ to both sides of (25) yields

$$\begin{aligned}&\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \|\Pi'_{X|T}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + (\tilde{f} - f^*)\|_2^2 + 2\lambda_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 + \lambda_2\|\tilde{f}\|_K^2 \\ &\le 3\delta_0 R^2 + 4\lambda_1\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}^*_S\|_1 + \lambda_2\|f^*\|_K^2 \le \frac{9}{2}\delta_0 R^2 + \frac{\|X'_T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2}{4}, \end{aligned} \tag{29}$$

where we use inequality (26) and assumptions $\frac{2\lambda_1^2 s^*}{\Lambda_{\min}^2} \le \delta_0 R^2$, $\lambda_2\|f^*\|_K^2 \le \delta_0 R^2$. Therefore, we have

$$\lambda_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \le \frac{9}{4}\delta_0 R^2,$$

and

$$\frac{\lambda_1\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_1}{R\sqrt{\delta_0/2}} \leq \frac{9}{4}\sqrt{\delta_0/2}R. \tag{30}$$

Combining (28) and (30) yields

$$\Omega_{R,\lambda}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \tilde{f}-f^*) \leq (10\delta_0)^{1/2}R + \frac{9}{4}\sqrt{\delta_0/2}R \leq \frac{R}{2},$$

as long as the small constant $\delta_0$ satisfies $\sqrt{\delta_0}\big[\sqrt{10}+\frac{9\sqrt{2}}{8}\big] \leq \frac{1}{2}$. Finally, our desired result follows from the following equality:

$$\Omega_{R,\lambda}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*, \tilde{f}-f^*) = s\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f) = \frac{R\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f)}{R+\Omega_{R,\lambda}(\widehat{\Delta}_{\boldsymbol{\beta}}, \widehat{\Delta}_f)}.$$

This completes the proof of Lemma 6. ■

At this point, if the event $\mathcal{T}(\delta_0, R)$ holds with high probability, the result of lemma 6 implies the convergence of the local estimator $(\hat{\boldsymbol{\beta}}, \hat{f})$ by choosing a proper small quantity of $R$. However, this derived convergence rate is rough and precisely the rate of the parametric estimator depends on that of the nonparametric estimator, which is suboptimal in the literature. In the appendix, we will show that $\mathcal{T}(\delta_0, R)$ holds with high probability under appropriate choices of regularization parameters.

### 5.3 Proof for Averaging Estimator

To derive the estimation error of our averaging parametric estimator, we decompose the total error into three parts: the first part characterizes the estimation error of the local estimator and the error of inverse matrix approximation, the second part reflects the approximation error of the nonparametric component in the RKHS, and the third part considers the total noise.

**Proof for Theorem 1**. Recall that the averaged parametric estimator $\bar{\boldsymbol{\beta}}$ defined on all the subsample is given by

$$\bar{\boldsymbol{\beta}} = \frac{1}{m}\sum_{l=1}^{m}\big[\hat{\boldsymbol{\beta}}^{(l)} + \frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I}-\mathbb{A}_u^{(l)}(\lambda_2))(\mathbf{Y}^{(l)}-\mathbb{X}^{(l)}\hat{\boldsymbol{\beta}}^{(l)})\big], \tag{31}$$

where $\hat{\boldsymbol{\beta}}^{(l)}$ is any Lasso estimator generated by minimizing $\mathcal{Q}^{(l)}(\boldsymbol{\beta})$ in (5) on the $l$-th subsample.

First, substituting the partially linear model into (31), we get

$$\begin{aligned}\bar{\boldsymbol{\beta}} = &\frac{1}{m}\sum_{l=1}^{m}\big[\hat{\boldsymbol{\beta}}^{(l)} - \frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I}-\mathbb{A}_u^{(l)}(\lambda_2))\mathbb{X}^{(l)}(\hat{\boldsymbol{\beta}}^{(l)}-\boldsymbol{\beta}^*)\big]\\ &+\frac{1}{m}\sum_{l=1}^{m}\big[\frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I}-\mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\big] + \frac{1}{m}\sum_{l=1}^{m}\big[\frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I}-\mathbb{A}_u^{(l)})\boldsymbol{\epsilon}^{(l)}\big].\end{aligned}$$

Subtracting $\boldsymbol{\beta}^*$ on both sides of the above equation, we obtain

$$\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leq \Omega_1 + \Omega_2 + \Omega_3, \tag{32}$$

where

$$\Omega_1 = \frac{1}{m}\sum_{l=1}^{m} \Big\| \big(\mathbb{I} - \frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2))\mathbb{X}^{(l)}\big)(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\Big\|_\infty$$

and

$$\Omega_2 = \Big\| \frac{1}{m}\sum_{l=1}^{m} \frac{1}{n}\big[\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\big]\Big\|_\infty, \Omega_3 = \frac{1}{N}\Big\| \sum_{l=1}^{m} \big[\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})\boldsymbol{\epsilon}^{(l)}\big]\Big\|_\infty.$$

We first consider the term $\Omega_1$. For any $l \in [m]$, it is straightforward to see each term in the sum is bounded by

$$\begin{aligned}
&\big\| \big(\mathbb{I} - \frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2))\mathbb{X}^{(l)}\big)(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\big\|_\infty \\
&\leq \big\| \big(\mathbb{I} - \Xi^{-1}\tilde{\Sigma}_l\big)(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\big\|_\infty + \big\| \big(\Xi^{-1} - \hat{\Theta}^{(l)}\big)(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\big\|_\infty \\
&\leq \max_{j\in[p]} \big\|\Xi_{j,\cdot}^{-1}\tilde{\Sigma}_l - \mathbf{e}_j\big\|_\infty \big\|\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*\big\|_1 + \max_{j\in[p]} \big\| \big(\Xi_j^{-1} - \hat{\Theta}_j^{(l)}\big)\big\|_1 \big\|\tilde{\Sigma}_l(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\big\|_\infty.
\end{aligned}$$

By Assumption E, $\max_{j\in[p]} \big\|\Xi_{j,\cdot}^{-1}\tilde{\Sigma}_l - \mathbf{e}_j\big\|_\infty = O_p(\sqrt{\log p/n})$ and $\max_{j\in[p]} \big\|\Xi_j^{-1} - \hat{\Theta}_j^{(l)}\big\|_1 = O_p(s_{max}\sqrt{\log p/n})$. By Theorem 3, we have $\big\|\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*\big\|_1 = O_p(s^*\sqrt{\log p/n})$. Also by the optimality conditions of the local estimators (5), we get $\big\|\tilde{\Sigma}_l(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta}^*)\big\|_\infty = O_p(\lambda_1)$. We put all the pieces together to obtain

$$\Omega_1 = (s^* + s_{max})O_p\Big(\frac{\log p}{n}\Big), \tag{33}$$

provided that all the conditions in Theorem 3 are satisfied.

Next, we provide an upper bound for $\Omega_2$. It is easy to see that $\Omega_2$ is bounded uniformly by $\Big\|\frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\Big\|_2$ for all $l \in [m]$. We notice that

$$\begin{aligned}
\Big\|\frac{1}{n}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\Big\|_2^2 &= \frac{1}{n^2}\big((\mathbf{f}^*)^{(l)}\big)'(\mathbb{I} - \mathbb{A}_u^{(l)})\mathbb{X}^{(l)}\hat{\Theta}^{(l)}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)} \\
&= \frac{1}{n^2} tr\Big((\mathbb{I} - \mathbb{A}_u^{(l)})\mathbb{X}^{(l)}\hat{\Theta}^{(l)}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'(\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\big((\mathbf{f}^*)^{(l)}\big)'\Big) \\
&\leq \frac{1}{n^2}\lambda_{\max}\big((\mathbb{I} - \mathbb{A}_u^{(l)})\mathbb{X}^{(l)}\hat{\Theta}^{(l)}\hat{\Theta}^{(l)}(\mathbb{X}^{(l)})'\big) tr\big((\mathbb{I} - \mathbb{A}_u^{(l)})(\mathbf{f}^*)^{(l)}\big((\mathbf{f}^*)^{(l)}\big)'\big) \\
&\leq \frac{1}{n}\lambda_{\max}(\hat{\Theta}^{(l)})\|(\mathbb{I} - \mathbb{A}_u^{(l)})^{1/2}(\mathbf{f}^*)^{(l)}\|_2^2,
\end{aligned} \tag{34}$$

where the second equality follows from $tr(AB) = tr(BA)$ and the first inequality uses $tr(AB) \leq \lambda_{\max}(A)tr(B)$ for any $A \succeq 0$ and $B$. Next, we define an map $S_t : \mathcal{H}_K \to \mathbb{R}^n$

by $S_t(f) = (f(T_1), ..., f(T_n))$. Denote by $S_t^*$ the adjoint map of $S_t$, satisfying $\langle S_t f, V\rangle_2 = \langle f, S_t^* V\rangle_K$ for any $f \in \mathcal{H}_K$ and $V \in \mathbb{R}^n$. Thus, we have

$$\begin{aligned}\|(\mathbb{I} - \mathbb{A}_u^{(l)})^{1/2}(\mathbf{f}^*)^{(l)}\|_2^2 &= \Big\langle S_t f^*, (\mathbb{I} - \mathbb{A}_u^{(l)}) S_t f^* \Big\rangle_2 = \Big\langle f^*, S_t^*(\mathbb{I} - \mathbb{A}_u^{(l)}) S_t f^* \Big\rangle_K \\ &\le \|S_t^*(\mathbb{I} - \mathbb{A}_u^{(l)}) S_t\|_{op} \|f^*\|_K^2,\end{aligned}$$

where $\|\cdot\|_{op}$ denotes the operator norm on $\mathcal{H}_K$. Moreover, by the property of adjoint map in Lemma 12, we know that

$$\begin{aligned}\|S_t^*(\mathbb{I} - \mathbb{A}_u^{(l)}) S_t\|_{op} &= \|(\mathbb{I} - \mathbb{A}_u^{(l)})^{1/2} S_t S_t^* (\mathbb{I} - \mathbb{A}_u^{(l)})^{1/2}\|_{op} \\ &= \|S_t S_t^* (\mathbb{I} - \mathbb{A}_u^{(l)})\|_{op} \\ &\le \lambda_2,\end{aligned}$$

where we use the fact $S_t S_t^* = \mathbb{K}_t^{(l)}$. Thus, this together with the above results immediately yields

$$\Omega_2 \le \|f^*\|_K \sqrt{\lambda_{\max}(\hat{\Theta}^{(l)})} \sqrt{\frac{\lambda_2}{n}}. \tag{35}$$

It remains to quantify $\Omega_3$. Note that the $j$-th element of $\Omega_3$ has the form

$$\Omega_{3j} = \frac{1}{N} \Big| \sum_{l=1}^{m} \sum_{i=1}^{n} w_{ij}^{(l)} \epsilon_i^{(l)} \Big|, \ j \in [p],$$

where $w_{\cdot j}^{(l)} := e_j^T W^{(l)}$ denotes the $j$-th row of $W^{(l)} := \hat{\Theta}^{(l)} (\mathbb{X}^{(l)})' (\mathbb{I} - \mathbb{A}_u^{(l)})$, $j \in [p]$. Since $\epsilon_i$ is independent of covariates $(X_i, T_i)$ by Assumption A and $\boldsymbol{\epsilon}^{(l)}$ for all $l \in [m]$ are not overlapping by splitting sample independently, $\Omega_{3j}$ is the sum of zero-mean i.i.d. Gaussian random variables conditional on $(X_i, T_i)$. Therefore, applying the Hoeffding-type inequality in Lemma 10 implies

$$\mathbb{P}[|\Omega_{3j}| > t | (X, T)] \le e \cdot \exp\Big( - \frac{cN^2 t^2}{2m \max_l \|w_{\cdot j}^{(l)}\|_2^2} \Big), \ \ \forall\, t > 0, \text{and for all } j \in [p]. \tag{36}$$

To continue, we need to provide an upper bound for $\max_{l,j} \|w_{\cdot j}^{(l)}\|_2^2$. In fact, a direct calculation yields that

$$\begin{aligned}\|w_{\cdot j}^{(l)}\|_2^2 &= e_j^T \hat{\Theta}^{(l)} (\mathbb{X}^{(l)})' (\mathbb{I} - \mathbb{A}_u^{(l)})^2 \mathbb{X}^{(l)} \hat{\Theta}^{(l)} e_j \\ &\le e_j^T \hat{\Theta}^{(l)} (\mathbb{X}^{(l)})' (\mathbb{I} - \mathbb{A}_u^{(l)}) \mathbb{X}^{(l)} \hat{\Theta}^{(l)} e_j \\ &\le n (\hat{\Theta}^{(l)})_{jj}.\end{aligned}$$

Then, with probability at least $1 - p^{-1}$, it follows from (36) that

$$\|\Omega_3\|_\infty \le c \sqrt{\max_j \{(\hat{\Theta}^{(l)})_{jj}\}} \sqrt{\frac{\log(3p)}{N}} \tag{37}$$

for some constant c.

Consequently, substituting (33), (35) and (37) into (32), we can complete the proof of Theorem 1. □

## 6. Simulations

We illustrate the performances of the distributed estimators via simulations. We generate the data from the model (1), where $\boldsymbol{\beta}^* = (1, 2, -1, 0.5, -2, 0, \ldots, 0)$ and $\epsilon_i \sim N(0, 4)$. We then generate a vector $Z_i$ in $\mathbb{R}^p$ from a mean-zero multivariate Gaussian distribution with correlations $\mathrm{Cov}(Z_{ij}, Z_{ij'}) = 0.3^{|j-j'|}$, $1 \le j, j' \le p$ and then set $T_i = \Phi(Z_{i1})$ and $X_{ij} = Z_{ij}, j = 2, \ldots, p$, where $\Phi$ is the cumulative distribution function of the standard normal distribution so that $T_i \in (0, 1)$. The nonparametric function is $f^*(t) = 5\sin(2\pi t)/(2 - \sin(2\pi t))$ and the RKHS is chosen to be the 3rd order Sobolev space. We select the tuning parameters in the penalties by 5-fold cross-validation in each local machine.

We compute the centralized estimator (CEN) for $\boldsymbol{\beta}$, the naive aggregated estimator without using bias correction (NAI) and the proposed aggregated estimator after bias correction (ABC). The accuracy of the estimators is assessed by $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty$.

First, we set $N = 2000$, $m = 1, 10$ ($m = 1$ is the centralized estimator) and $p = 100, 200, 400, 800, 1600$. We generate 200 data sets for each setting. Figure 1 shows the average errors of the centralized estimator (black) and those of the distributed estimators with $m = 10$. It is seen the performance becomes worse with dimension as expected. The proposed aggregated estimator after bias correction (ABC) performs better than the naive aggregated estimator without using bias correction (NAI) for all dimensions.

In the second set of simulations, we vary $m = 1, 5, 10, 20, 25$ while fixing $N = 2000$ or 8000, and $p = 1000$. The performances generally deteriorate with the increase of $m$. Again, in terms of $l_\infty$ error, ABC is better than NAI.

In the final set of simulations, we consider larger sample sizes $N = 2000, 4000, 6000, 8000, 10000$, with $p = 1000$, and fix the size of the sub-sample in each local machine to be $n = 200$. It is seen that ABC has errors decreasing with total sample size, while the naive aggregated estimator NAI has larger errors.

The simulations are carried out on the computational cluster Katana in the University of New South Wales. For the first set of simulations with $p = 1600$ for example, the central estimators require about 8 hours to finish all 200 repetitions, while the distributed estimator with $m = 10$ requires about 1.5 hours.

## 7. Conclusions

Although distributed estimation or distributed learning has been studied well for linear models and fully nonparametric models, to date partial linear models have been rarely studied under the distributed setting. The latter encounters additional difficulty even in contrast to the centralized method on the entire data. As shown in the literature, the linear part in PLM can be estimated with oracle rates as if the nonparametric component were known, even though the rate for estimating the nonparametric component is slower than the oracle rate for the linear part. By contrast, to derive non-asymptotic oracle rates for the averaging parametric estimator, the smoothness of kernel-based nonparametric function significantly affects the number of data partitions. To handle this problem, we prove the oracle rate for the linear part with a novel technical proof, thereby yielding the minimax optimal rate of the parametric estimator in some senses.

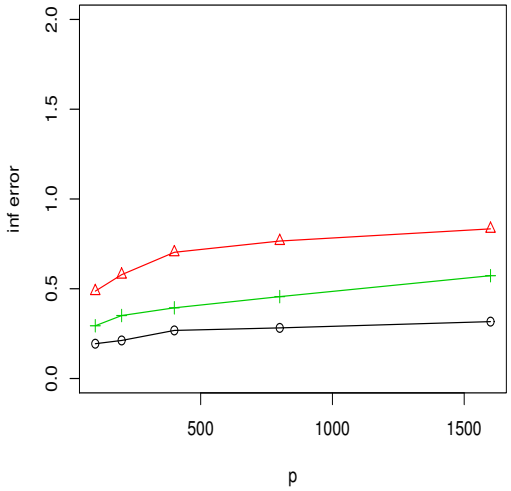


Figure 1: The $l_\infty$ errors of estimates with changing dimension $p$. $\circ$(black): centralized estimator (CEN); $\triangle$(red): naive aggregated estimator (NAI); $+$(green): the aggregated estimator after debiasing (ABC).

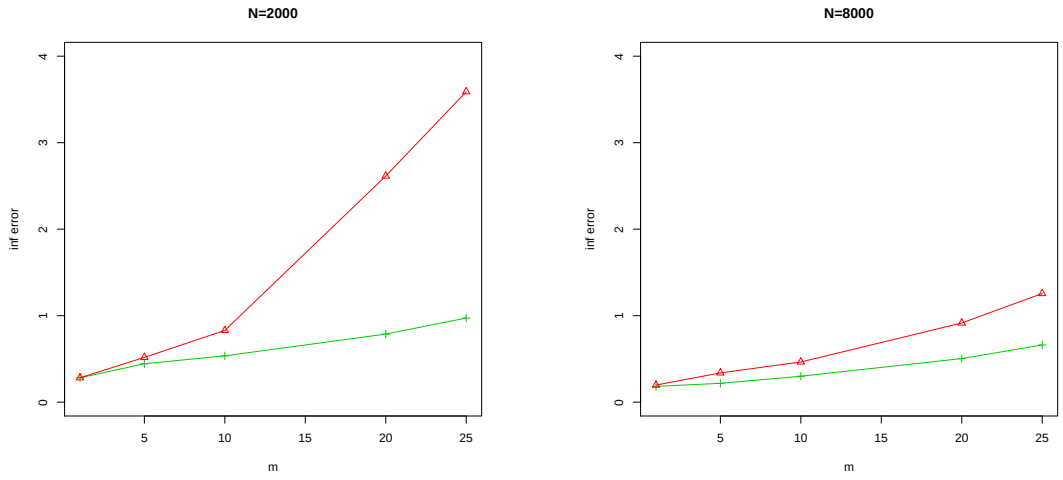


Figure 2: The $l_\infty$ errors of estimates with $m \in \{1, 5, 10, 20, 25\}$ ($m = 1$ represents the centralized estimator). $\triangle$(red): naive aggregated estimator (NAI); $+$(green): the aggregated estimator after debiasing (ABC). Left panel: $n = 2000$, right panel: $n = 8000$.

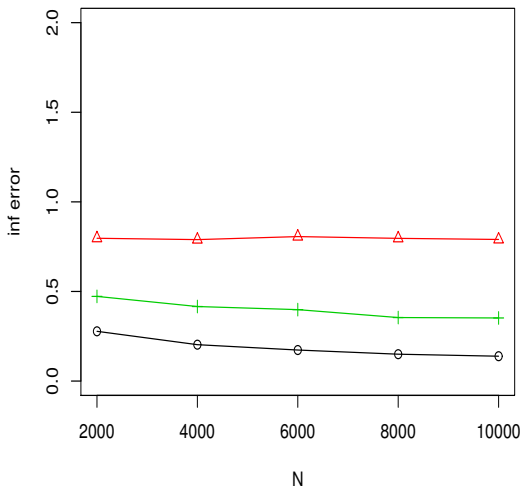


Figure 3: The $l_\infty$ errors of estimates with $p = 1000$ and $N \in \{2000, 4000, 6000, 8000, 10000\}$. $\circ$(black): centralized estimator (CEN); $\triangle$(red): naive aggregated estimator (NAI); +(green): the aggregated estimator after debiasing (ABC).

On the other hand, the classical doubly regularized approach for estimating the sparse PLM heavily leads to estimation bias due to the two convex penalty terms. Hence, how to reduce bias is a critical issue to improve inference efficiency for the corresponding distributed estimation. We transform the proposed estimation into a Lasso-type optimization only containing parametric coefficients, and then propose a new debiased distributed estimation for the sparse PLM under high dimensional settings, showing comparable numerical performance through several simulation experiments.

We remark that forming $\hat{\Theta}^{(l)}$ is $p$-times as expensive as solving the local Lasso problem, which is the most expensive step of evaluating the averaging estimator. To this end, we could consider an estimator only using a common $\hat{\Theta}$ for all the local estimators in the following way. To reduce computational cost, we assign the task of computing $p/m$ rows of $\hat{\Theta}$ to each local machine. Then each machine sends $\frac{p}{m}$ rows of $\hat{\Theta}$ computed by (7) to the central server, as well as sending $\hat{\boldsymbol{\beta}}^{(l)}$ and $(\mathbb{X}^{(l)})'\big(\mathbb{I} - \mathbb{A}_u^{(l)}(\lambda_2)\big)(\mathbf{Y}^{(l)} - \mathbb{X}^{(l)}\hat{\boldsymbol{\beta}}^{(l)})$. Here we use different $\hat{\Theta}^{(l)}$ for different machines merely for convenience of presentation and implementation.

## Acknowledgments

We are grateful to two referees and the associate editor for valuable comments and constructive suggestions. The research of Shaogao Lv is supported by National Natural Science Foundation of China (Grant No.11871277 and 11829101). The research of Heng Lian is supported by Project 11871411 from NSFC and CityU Shenzhen Research Institute, and by Hong Kong RGC general research fund 11301718, 11300519, 11300721, and 11311822.

## Appendix A.

### Appendix A. Concentration Inequalities and Complexity Bounds

In this appendix we list several technical lemmas.

**Lemma 7** *(Talagrand's Concentration Inequality) Let $\mathcal{G}$ be a function class on $\mathcal{Z}$ that is separable with respect to $\infty$-norm, and $\{z_i\}_{i=1}^n$ be i.i.d. random variables with values in $\mathcal{Z}$. Furthermore, let $B > 0$ and $U \geq 0$ be $B := \sup_{g\in\mathcal{G}} \mathbb{E}[(g - \mathbb{E}[g])^2]$ and $U := \sup_{g\in\mathcal{G}} \|g\|_\infty$, then there exists a universal constant $N_0$ such that, for $\mathbf{Z} = \sup_{g\in\mathcal{G}} \big|\frac{1}{n}\sum_{i=1}^n g(z_i) - \mathbb{E}[g]\big|$, we have*

$$\mathbb{P}\Big(\mathbf{Z} \geq N_0\Big[\mathbb{E}[\mathbf{Z}] + \sqrt{\frac{Br}{n}} + \frac{Ur}{n}\Big]\Big) \leq e^{-r}, \quad \forall\, r > 0.$$

We denote by $\{\sigma_i\}_{i=1}^n$ the Rademacher random variables that are an i.i.d. random variables taking values in $\{-1, +1\}$ with probability $1/2$. Recall that, for a set of measurable functions $\mathcal{F}$ that is separable with respect to $\infty$-norm, the Rademacher complexity $\mathrm{R}_n(\mathcal{F}, \Phi(f) \leq r) := \mathbb{E}_{\sigma,u}\Big[\sup_{f\in\mathcal{F},\Phi(f)\leq r} \frac{1}{n}\big|\sum_{i=1}^n \sigma_i f(u_i)\big|\Big]$ of $\mathcal{F}$ controls the supremum of discrepancy between the empirical and population means of all functions $f \in \mathcal{F}$ (see Lemma 2.3.3 of van der Vaart and Wellner (1996)):

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \big|\frac{1}{n}\sum_{i=1}^n (f(u_i) - \mathbb{E}[f])\big|\Big] \leq 2\mathrm{R}_n(\mathcal{F}, \mathbb{E}[f^2] \leq r). \tag{38}$$

**Lemma 8** *Let $\mathcal{F}$ be a class of functions with ranges in $[a, b]$ and there are some functional $\Phi : \mathcal{F} \to \mathbb{R}^+$ such that for every $f \in \mathcal{F}$, $Var[f] \leq \Phi(f) \leq B\mathbb{E}[f]$. Let $\psi$ be a sub-root function and $r^*$ be the fixed point of $\psi$. Furthermore, assume that $\psi$ satisfies, for any $r \geq r^*$, $\psi(r) \geq B\mathrm{R}_n(\mathcal{F}, \Phi(f) \leq r)$. Then, with $c_1 = 704$ and $c_2 = 26$, for any $L > 1$ and every $r > 0$, with probability at least $1 - e^{-r}$,*

$$\frac{L}{L-1}\Big(\mathbb{E}[f] - \frac{1}{n}\sum_{i=1}^n f(u_i)\Big) \leq \frac{1}{L-1}\mathbb{E}[f] + \frac{c_1 L}{B} r^* + \frac{r(11(b-a) + c_2 LB)}{n}, \ \ \forall\, f \in \mathcal{F}.$$

The above concentration inequality can be viewed as a simple version of Theorem 3.3 in Bartlett et al. (2005).

The following result was proved in Mendelson (2002). There is an interesting finding that the upper bound of the Rademacher complexity in the RKHS is independent of the dimension.

**Lemma 9** *Suppose that the general kernel $K$ is bounded uniformly by $\kappa$, then there holds*

$$\mathrm{R}_n(\mathbb{B}_K(1)) \leq \sqrt{\frac{2\kappa}{n}}.$$

*Moreover, there also holds*

$$\mathrm{R}_n(f \in \mathbb{B}_K(1), \|f\|_2 \leq r) \leq \mathcal{Q}_n(r).$$

For any $g = X'\boldsymbol{\beta} + f(T) \in \mathcal{G}(R)$, we easily check that $\|f\|_K \leq \frac{R}{\sqrt{\lambda_2}}$. Moreover, the traingle inequality is applied to obtain $\|g\|_2^2 = \|X_T'\boldsymbol{\beta}\|_2^2 + \|\Pi_{X|T}'\boldsymbol{\beta} + f(T)\|_2^2 \leq R^2$. This together with Assumption C implies that $\|\boldsymbol{\beta}\|_2 \leq R/\Lambda_{\min}$. Furthermore, the triangle inequality is used to imply

$$\|f\|_2 \leq \|g\|_2 + \|X'\boldsymbol{\beta}\|_2 \leq \big(\Gamma_{\max}/\Lambda_{\min} + 1\big)R.$$

Based on this, we obtain

$$\mathbb{E}\Big( \sup_{g\in\mathcal{G}(R)} \Big|\frac{1}{n}\sum_{i=1}^n f(T_i)\sigma_i\Big|\Big) \leq \frac{R}{\sqrt{\lambda_2}}\mathrm{R}_n(f \in \mathbb{B}_K(1), \|f\|_2 \leq \sqrt{\lambda_2}) \leq \frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\sqrt{\lambda_2}}R. \tag{39}$$

The following lemma belongs to one of large deviation inequalities for sums of independent sub-Gaussian random variables, and can be found in Proposition 5.10 in Vershynin (2011). The sub-Guassian norm of $X$ is defined by $\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$.

**Lemma 10** *(Hoffding-type inequality). Let $Z_1, ..., Z_n$ be independent centered sub-gaussian random variables, and let $D = \max_i \|Z_i\|_{\psi_2}$. Then for every $a = (a_1, ..., a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n a_i Z_i\Big| \geq t\Big) \leq e\exp\Big(-\frac{ct^2}{D^2\|a\|_2^2}\Big),$$

*where $e$ is some universal constant.*

We introduce the event involving $\nu_n$ defined in the text previously:

$$\mathcal{E}(\nu_n) = \Big\{\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(T_i)\Big)\Big| \leq \nu_n^2\|f\|_K + \nu_n\|f\|_2, \quad f \in \mathcal{H}_K\Big\}.$$

By simplifying Theorem 10 for multi-kernel regression problems in the supplementary material of Suzuki and Sugiyama (2013), one shows that the event $\mathcal{E}(\nu_n)$ occurs with high probability, stated as follows.

**Lemma 11** *Suppose that $\epsilon_i$'s are independent Gaussian variables. Under the Supernorm Assumption, there exist two constants $c_1$, $c_2$ such that*

$$\mathbb{P}[\mathcal{E}(\nu_n)] \geq 1 - c_1\exp(-c_2 n\nu_n^2).$$

In the end, we list the classical conclusion on the adjoint operators in Hilbert spaces (see Chapter 8 in Rudin (1991)).

**Lemma 12** *Let $H_1, H_2$ be two Hilbert spaces, and $A$ is a linear and bounded operator from $H_1$ to $H_2$, with its adjoint operator $A^*$. Then, $\|A\| = \|A^*\| = \|AA^*\|^{1/2} = \|A^*A\|^{1/2}$.*

## Appendix B. Proof for $\mathcal{T}_1(\delta_0, R)$

**Lemma 13** *Suppose that Assumptions A-D hold, with $\lambda_2 \simeq R^2$ and $R^2 \leq \lambda_1 \leq 1$. For constants $\delta_1, \delta_1''$ and $\kappa_1$ with our suitable choices, we set $\lambda_0 := \sqrt{\log(2p)/n}$ and*

$$\delta_1\lambda_1 \geq \lambda_0,\ \sqrt{n}\lambda_1 \geq 1,\ \lambda_2 \geq \kappa_1 n^{-\frac{1}{\tau+1}}.$$

*Then we conclude*

$$\sup_{g\in\mathcal{G}(R)} \big|\|g\|_n^2 - \|g\|^2\big| \leq \delta_0 R^2$$

*with probability at least* $1 - \exp[-n(\delta_1'')^2\lambda_2]$.

**Proof** In order to verify all the conditions of Lemma 7, we denote $\boldsymbol{Z} := \sup_{g\in\mathcal{G}(R)} \big|\|g\|_n^2 - \|g\|^2\big|$ with $\mathcal{G}(R)$ and $Z := (X, T)$. By a direct computation, we have

$$\|g^2\|_\infty = \big\|\big(X'\boldsymbol{\beta} + f(T)\big)^2\big\|_\infty \leq 2\|f\|_K^2 + 2C_0^2\|\boldsymbol{\beta}\|_1^2.$$

Note that for $g \in \mathcal{G}(R)$, it follows that

$$\|\boldsymbol{\beta}\|_1 \leq \frac{\sqrt{\delta_0/2}R^2}{\lambda_1},\ \text{and}\ \|f\|_K^2 \leq \frac{R^2}{\lambda_2},$$

which implies that

$$\|g^2\|_\infty \leq 2\frac{R^2}{\lambda_2} + \delta_0 C_0^2\frac{R^4}{\lambda_1^2} \leq 2/c(\delta_0) + \delta_0 C_0^2,$$

due to the assumption $\lambda_2 \geq c(\delta_0)R^2$ and $R^2 \leq \lambda_1 \leq 1$. Setting $\tilde{C} = 2/c(\delta_0) + \delta_0 C_0^2$, for any $g \in \mathcal{G}(R)$, we further have

$$var(g^2) \leq \mathbb{E}[g^4] \leq \|g^2\|_\infty \mathbb{E}[g^2] \leq \tilde{C}R^2. \tag{40}$$

We now need to provide an upper bound of $\mathbb{E}[\boldsymbol{Z}]$. Let $\{\sigma_i\}_{i=1}^n$ be a Rademacher sequence independent of $\{(X_i, T_i)\}_{i=1}^n$. By symmetrization [see e.g. van der Vaart and Wellner (1996)], we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{Z}] \leq\ & 2\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n g_i^2\sigma_i\Big|\Big) \leq 2\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n \langle f^2(T_i)\sigma_i\Big|\Big) \\ &+ 2\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n (X_i'\boldsymbol{\beta})^2\sigma_i\Big|\Big) + 4\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n (X_i'\boldsymbol{\beta})f(T_i)\sigma_i\Big|\Big).\end{aligned}$$

In the following, we bound the above three quantities respectively. Note that for $f(T) + X'\boldsymbol{\beta} \in \mathcal{G}(R)$, Condition B leads to

$$\|X'\boldsymbol{\beta}\|_\infty \leq C_0\frac{\sqrt{\delta_0/2}R^2}{\lambda_1} \leq C_0\sqrt{\frac{\delta_0}{2}},$$

where we use the assumption that $R^2 \leq \lambda_1$. By the contraction inequality of Rademacher complexity [see Ledoux and Talagrand (1991)], we get

$$\mathbb{E}\Big(\sup_{\boldsymbol{\beta}, f(T)+Z^T\boldsymbol{\beta}\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n (X_i'\boldsymbol{\beta})^2\sigma_i\Big|\Big) \leq 4C_0\sqrt{\frac{\delta_0}{2}}\mathbb{E}\Big(\sup_{\boldsymbol{\beta}, f(T)+X'\boldsymbol{\beta}\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n (X_i'\boldsymbol{\beta})\sigma_i\Big|\Big).$$

Moreover, we know that

$$\mathbb{E}\Big(\sup_{\boldsymbol{\beta},f(T)+X'\boldsymbol{\beta}\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i'\boldsymbol{\beta})\sigma_i\Big|\Big)\le\mathbb{E}\Big(\sup_{\boldsymbol{\beta},f(T)+Z^T\boldsymbol{\beta}\in\mathcal{G}(R)}\Big\|\frac{1}{n}\sum_{i=1}^{n}X_i\sigma_i\Big\|_\infty\|\boldsymbol{\beta}\|_1\Big)$$
$$\le\frac{\sqrt{\delta_0/2}R^2}{\lambda_1}\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^{n}X_i\sigma_i\Big\|_\infty\le C_0\frac{\sqrt{\log(2p)/n}R^2}{\sqrt{1/\delta_0}\lambda_1}\le\big(\delta_1C_0\sqrt{\delta_0}\big)R^2,$$

where the first inequality follows from the Cauchy-Schwarz inequality, the classical concentration result $\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^{n}X_i\sigma_i\Big\|_\infty\le C_0\sqrt{2\log(2p)/n}$ in the third one and the assumption $\delta_1\lambda_1\ge\sqrt{\log(2p)/n}$ for the last inequality. Hence, combining with the above two inequalities yields

$$\mathbb{E}\Big(\sup_{\boldsymbol{\beta},f(T)+X'\boldsymbol{\beta}\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i'\boldsymbol{\beta})^2\sigma_i\Big|\Big)\le2\delta_1C_0^2\delta_0R^2. \tag{41}$$

At this point, we still require a tight bound on $\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}f(T_i)^2\sigma_i\Big|\Big)$. As stated above, it is shown that $|f(T)|\le\|f\|_K\le\frac{R}{\sqrt{\lambda_2}}$. By the contraction property of Rademacher sequences again, we have

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}f(T_i)^2\sigma_i\Big|\Big)\le2\frac{R}{\sqrt{\lambda_2}}\cdot\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}f(T_i)\sigma_i\Big|\Big)$$
$$\le2\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_2}R^2, \tag{42}$$

which follows from the obtained result in (39) in the Appendix. Similarly, we also have

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}(X_i'\boldsymbol{\beta})f(T_i)\sigma_i\Big|\Big)\le\mathbb{E}\sup_{g\in\mathcal{G}(R)}\Big\|\frac{1}{n}\sum_{i=1}^{n}X_if(T_i)\sigma_i\Big\|_\infty\|\boldsymbol{\beta}\|_1 \tag{43}$$
$$\le\sqrt{\delta_0/2}\frac{R^2}{\lambda_1}\mathbb{E}\max_{1\le j\le p}\sup_{f,X'\boldsymbol{\beta}+f(T)\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}X_{ij}f(T_i)\sigma_i\Big|$$
$$\le C_0\sqrt{\delta_0/2}\frac{R^2}{\lambda_1}\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}f(T_i)\sigma_i\Big|\Big)$$
$$\le C_0\sqrt{\delta_0/2}\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_1\sqrt{\lambda_2}}R\cdot R^2 \tag{44}$$

where the third inequality follows from the contraction property of Rademacher complexity, and the last inequality follows from (39) below. Along the lines of (41), (42) and (43), we get

$$\mathbb{E}[\boldsymbol{Z}]\le\Big(2\delta_1C_0^2\delta_0+2\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_2}+C_0\sqrt{\delta_0/2}\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_1\sqrt{\lambda_2}}R\Big)R^2.$$

Therefore, by the concentration theorem in Lemma 3, we have

$$\mathbb{P}\Big(\boldsymbol{Z}\ge\tilde{D}R^2+N_0\sqrt{\tilde{C}}\frac{R\sqrt{t}}{\sqrt{n}}+\frac{2N_0\tilde{C}t}{3n}\Big)\le\exp[-t],\quad\forall\,t>0,$$

where $\tilde{D} := 2N_0\delta_1 C_0^2\delta_0 + 2N_0\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_2} + N_0C_0\sqrt{\delta_0/2}\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_1\sqrt{\lambda_2}}R$. Note that, by the spectral assumption (Assumption D(i)), it is easy to check that $\mathcal{Q}_n(\sqrt{\lambda_2}) = O(\frac{\lambda_2^{(1-\tau)/2}}{\sqrt{n}})$ and thus $\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_2} = O(\frac{1}{\sqrt{n\lambda_2^{(1+\tau)}}})) = O(1/\sqrt{\kappa_1})$ following the assumption $n\lambda_2^{(1+\tau)} \geq \kappa_1$. Similarly, we also have $\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\lambda_1\sqrt{\lambda_2}}R = O(\sqrt{\frac{\lambda_2^{(1-\tau)}}{\kappa_1}})$ based on $R^2 \leq \lambda_1$ and $\sqrt{n}\lambda_1 \geq 1$. We now take $t = n(\delta_1'')^2\lambda_2$ and assume that $\lambda_2 \leq c_1R^2$ with some constant $c_1$. Taking $\delta_1$ and $\delta_1''$ small enough but $\kappa_1$ large enough, such that

$$\tilde{D} + N_0\sqrt{\tilde{C}}\delta_1'' + 2\sqrt{c_1}\tilde{C}D(\delta_1'')^2 + \frac{2}{3}c_1N_0\tilde{C}(\delta_1'')^2 \leq \delta_0,$$

then we have

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}(R)}\big|\|g\|_n^2 - \|g\|^2\big| \geq \delta_0R^2\Big) \leq \exp[-n(\delta_1'')^2\lambda_2].$$

■

## Appendix C. Proof for $\mathcal{T}_2(\delta_0, R)$

To verify that the event $\mathcal{T}_2(\delta_0, R)$ occurs with high probability, we make use of an upper bound of the Gaussian process stated as follows.

**Lemma 14** *With the same conditions as Lemma 13, we have*

$$\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_ig(X_i,T_i)\Big| \leq \delta_0R^2,$$

*with probability at least* $1 - \exp[-n(\delta_1'')^2\lambda_2]$.

The Gaussian concentration inequality from Theorem 7.1 of Ledoux (2001) is a useful tool in our refined analysis, which provides tighter bounds than the general sub-Gaussian cases. In particular, the super-norm bounds of random variables are not needed, as opposed to Rademacher concentration inequality presented in Lemma 7.

**Lemma 15** *Let* $\mathbb{G} = \{G_t\}_{t\in T}$ *be a centered Gaussian process indexed by a countable set* $T$ *such that* $\sup_{t\in T}G_t < \infty$ *almost surely. Then*

$$\mathbb{P}\Big(\sup_{t\in T}G_t \geq \mathbb{E}[\sup_{t\in T}G_t] + \sqrt{r}\Big) \leq \exp(-\frac{r}{2\sigma^2}),$$

*where* $\sigma^2 = \sup_{t\in T}\mathbb{E}[G_t^2] < \infty$.

**Proof of Lemma 14.** Note that for any $g \in \mathcal{G}(R)$,

$$\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_ig(X_i,T_i)\Big| \leq \Big|\frac{1}{n}\sum_{i=1}^n\epsilon_if(T_i)\Big| + \sup_j\Big|\frac{1}{n}\sum_{i=1}^n\epsilon_iX_{ij}\Big|\|\boldsymbol{\beta}\|_1.$$

On one hand, we conclude from the conclusion in (39) below that

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(T_i)\Big|\Big)\leq\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\sqrt{\lambda_2}}R.$$

In addition, since $\epsilon_i$'s are standard Gaussian variables and $|X_{ij}|\leq C_0$ by Assumption B, Bernstein inequality is applied to yield

$$\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^{n}X_i\epsilon_i\Big\|_\infty\leq C_0\sqrt{\log(2p)/n}.$$

Thus, using similar arguments to (41) and (42) yields that

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i,T_i)\Big|&\leq\frac{\mathcal{Q}_n(\sqrt{\lambda_2})}{\sqrt{\lambda_2}}R+C_0\sqrt{\frac{\delta_0\log(2p)}{2n}}\frac{R^2}{\lambda_1}\\&\leq\sqrt{\frac{c_1}{\kappa_1}}R^2+C_0\delta_1\sqrt{\delta_0/2}R^2,\end{aligned}$$

where we follow the same augments as that in the last proof part of Lemma 13. Observe that $\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i,T_i)\Big|$ is a centered Gaussian process by Assumption A, and also check that $\sigma^2\leq\frac{1}{n}R^2$ in Lemma 15. Then, by the Gaussian concentration inequality with $r=2(\delta_1'')^2\lambda_2R^2$, we have

$$\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(X_i,T_i)\Big|\leq\sqrt{c_1}(1/\sqrt{\kappa_1}+\sqrt{2}\delta_1'')R^2+C_0\delta_1\sqrt{\delta_0/2}R^2.$$

As long as $\delta_1,\delta_1''$ are small sufficiently and $\kappa_1$ is properly large, we can obtain the desired result. □

## References


N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33:1497–1537, 2005.

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed estimation and inference with statistical guarantees. *The Annals of Statistics*, 46:1352–1382, 2018.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

H. Chen. Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, 16:136–146, 1988.

P. J. Green and B. S. Yandell. *Semi-parametric Generalized Linear Models*. Springer, New York, 1985.

W. Hardle and H. Liang. *Partially Linear Models*. Springer, New York, 2007.

N. E. Heckman. Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2):244–248, 1986.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

M. Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, Providence, RI, 2001.

J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor. Communication-efficient sparse regression: a one-shot approach. *Journal of Machine Learning Research*, 18:1–30, 2017.

H. Lian, X. Chen, and J. Yang. Identification of partially linear structure in additive models with an application to gene expression prediction from sequences. *Biometrics*, 68 (2):437–445, 2012.

H. Lian, K. Zhao, and S. Lv. Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data. *The Annals of Statistics*, 47 (5):2922–2949, 2019.

S. B. Lin, X. Guo, and D. X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

E. Mammen and S. van de Geer. Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, 25(3):1014–1035, 1997.

R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

S. Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43, 2002.

P. Müller and S. van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42:580–608, 2015.

X. Ni, H. H. Zhang, and D. W. Zhang. Automatic model selection for partially linear models. *Journal of Multivariate Analysis*, 100(9):2100–2111, 2009.

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

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

J. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

W. Rudin. *Functional Analysis*. McGraw-Hill, 1991.

P. Speckman. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):413–436, 1988.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the International Conference on Learning Theory*, pages 79–93, 2009.

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3):1381–1405, 2013.

S. van der Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2011.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

L. Wang, L. Xue, A. Qu, and H. Liang. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, 42(2):592–624, 2014.

Q. Wu. Bias corrected regularization kernel network and its applications. In *2017 International Joint Conference on Neural Networks*, pages 1072–1079, 2017.

H. L. Xie and J. Huang. Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, 37(2):673–696, 2009.

K. Yu, E. Mammen, and B. U. Park. Semi-parametric regression: efficiency gains from modeling the nonparametric part. *Bernoulli*, 17(2):736–748, 2011.

H. H. Zhang, G. Cheng, and Y. Liu. Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495): 1099–1112, 2011.

Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

T. Q. Zhao, G. Cheng, and H. Liu. A partially linear framework for massive heterogeneous data. *The Annals of statistics*, 44(4):1400, 2016.

Y. Zhu. Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients. *The Annals of Statistics*, 45(5):2274–2298, 2017.