A General Framework: Inference on High-Dimensional Nonparametric Models

Abstract

We propose a novel nonparametric model which naturally generalizes the sparse additive model and sparse varying coefficient model. Given covariate of interest X_1 and covariate of nuisance $\boldsymbol{X}_2$, our model assumes the mean function to be a sum of the product of a known function $\eta(\boldsymbol{X}_2)$ and the function of interest $f(X_1)$, and the nuisance function which can be estimated in the form of basis expansion. We propose to infer the function of interest $f(\cdot)$ using a two step plug-in approach, given an initial estimator of nuisance. We study the issue of bias propagation, which describes the dependence structure of bias when estimating the function of interest $f(\cdot)$ on the bias when estimating nuisance. We quantifies such issue by establishing the rate of convergence estimating f under the supremum norm, in terms of estimation/approximation rate of nuisance. We also propose a confidence band for f via Gaussian multiplier bootstrap, with its coverage probability converging to the nominal probability at the polynomial rate. Thorough numerical results on synthetic data are provided to demonstrate the efficacy of the theory.

1 Introduction

Nonparametric regression investigates the relationship between a target variable Y and many input variables $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ without imposing strong assumptions. Consider a model

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^d$ is a d-dimensional random vector in $\mathbb{R}^d$, ε is random error satisfying $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$, and Y is a target variable. When d is small, fitting a fully nonparametric model (1.1) is feasible (Wasserman, 2006). However, the interpretation of these models is challenging. When d is large, consistently fitting $f(\cdot)$ must rely on additional structural assumptions due to curse of dimensionality.

Among simple and powerful nonparametric models for dimensionality reduction, most important two of them are additive models (Friedman and Stuetzle, 1981; Stone, 1985; Hastie and Tibshirani, 1990) and varying coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999).

Specifically, in additive models we assume that

$$Y = \sum_{i=1}^{d} f_j(X_j) + \varepsilon, \quad \text{and} \quad \mathbb{E}_{X_j}[f_j(X_j)] = 0, \tag{1.2}$$

where $f_j(\cdot)$s are smooth univariate functions. Under the assumption that only s components are nonzero ($s \ll d$), significant progress has been made to understand additive models in high dimensions

(Lin et al., 2006; Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii et al., 2010; Kato, 2012).

Varying coefficient model assumes the following conditional linear structure

$$Y = \sum_{i=1}^{d} X_j \beta_j(Z) + \varepsilon, \tag{1.3}$$

for given covariates $(Z, X_1, \ldots, X_d)^{\mathrm{T}}$ and response variable Y. It is particularly appealing in longitudinal studies where it allows one to examine the extent to which covariates affect response over time. Under the assumption that only s components are nonzero ($s \ll d$), significant progress has been made to understand varying coefficient models in high dimensions (Lian, 2012; Fan et al., 2014) as well.

This paper proposes a novel model family which naturally generalizes both additive models and varying coefficient models. We assume

$$Y = f(X_1)\eta(\boldsymbol{X}_{\setminus 1}) + g(\boldsymbol{X}) + \varepsilon, \tag{1.4}$$

where $f(\cdot)$ is the function of interest, $\eta(\cdot)$ is a known function, and $g(\cdot)$ is nuisance. Once the function of interest is selected, we will observe such structure in additive model (1.2) and varying coefficient model (1.3). Under our model, we aim to estimate the marginal influence of the first input variable X_1 on the target Y. In particular, we are interested in estimating $f(\cdot)$ in a high dimensional setting and provide a confidence band for it.

The main results of our paper is twofold. First, we propose a unified framework for studying the estimation problem under both sparse additive model and sparse varying coefficient model. We propose a kernel estimator of f via a plug-in procedure, and establish a uniform rate of convergence of such estimator under our model. Intuitively, if the nuisance g in our model can be initially estimated with rate r_n, the function of interest f can then be estimated with a certain rate depending on r_n. We give a explicit form of this rate of convergence depending on r_n, and conditions that r_n, the rate of convergence of nuisance, needs to satisfy to yield a fast rate for estimating f. Therefore, our result serves as an interface to determine the estimation quality of function of interest f, given knowledge on the initial estimation quality of nuisance g.

Second, we propose one way to construct confidence bands for the function of interest f, and therefore provide a unified framework for studying inference problem under both models. To establish the validity of the proposed confidence bands, we develop three technical ingredients: (1) the analysis of suprema of a high dimensional empirical process that arises from kernel estimator, (2) a debasing method for the proposed estimator, and (3) the approximation analysis for the Gaussian multiplier bootstrap procedure. The supremum norm for our estimator is derived by applying results on the suprema of empirical processes (Koltchinskii, 2011; Van Der Vaart and Wellner, 1996; Bousquet, 2002). The de-biasing procedure for the kernel estimator extends the approach used in Ning and Liu (2014) to a inferring a nonparametric component of interest under a high dimensional semiparametric model. To prove the validity of the confidence band constructed by the Gaussian multiplier bootstrap, we generalize the method proposed in Chernozhukov et al. (2014a) and Chernozhukov et al. (2014b) to our high dimensional model.

1.1 Related Literature

Our work contributes to two different areas. For both areas, we made new methodological and technical contributions.

First, we contribute to the literature on high dimensional nonparametric estimation, which has recently seen a lot of activity. Lafferty and Wasserman (2008), Bertin et al. (2008), Comminges et al. (2012), Yang et al. (2015) and Linero (2018) study variable selection in a high dimensional nonparametric regression setting without assuming structural assumptions on $f(\cdot)$ beyond that it depends only on a subset of variables. A large number of papers have studied the sparse additive model in (1.2) (Sardy and Tseng, 2004; Lin et al., 2006; Avalos et al., 2007; Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii et al., 2010; Raskutti et al., 2012; Kato, 2012; Petersen et al., 2014; Rosasco et al., 2013; Lou et al., 2014; Wahl, 2014; Gregory et al., 2016; Yuan and Zhou, 2016; Guo and Zhang, 2019a; Yao and Zhang, 2020). In addition, Yao and Zhang (2020) proposes a multi-resolution group Lasso method without the knowledge of sparsity or smoothness, Yuan and Zhou (2016) study minimax optimal rates of convergence for estimation. Xu et al. (2014) study a high dimensional convex nonparametric regression. Dalalyan et al. (2014) study the compound model, which includes the additive model as a special case. Meier et al. (2009), Huang et al. (2010), Koltchinskii et al. (2010), Raskutti et al. (2012), and Kato (2012) develop estimation schemes mainly based on the basis approximation and sparsity-smoothness regularization.

Second, we contribute to a growing literature on high dimensional inference. Initial work on high dimensional statistics has focused on estimation and prediction (see, for example, Bühlmann and Van De Geer, 2011, for a recent overview) and much less work has been done on quantifying uncertainty, for example, hypothesis testing and confidence intervals. Recently, the focus has started to shift towards the latter problems. Initial work on construction of p-values in high dimensional models relied on correct inclusion of the relevant variables (Wasserman and Roeder, 2009; Meinshausen et al., 2009). Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) study stability selection procedure, which provides the family-wise error rate for any selection procedure. Hypothesis testing and confidence intervals for low dimensional parameters in high dimensional linear and generalized linear models are studied in Guo and Zhang (2019b), Cai and Guo (2020), Cai and Guo (2017), Belloni et al. (2018), Belloni et al. (2017), Belloni et al. (2014a), Belloni et al. (2013b), Van de Geer et al. (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2013), and Farrell (2013). These methods construct honest, uniformly valid confidence intervals and hypothesis test based on the ℓ_1 penalized estimator in the first stage. Similar results are obtained in the context of ℓ_1 penalized least absolute deviation and quantile regression (Belloni et al., 2014b, 2013a). Kozbur (2013) extends the approach developed in Belloni et al. (2014a) to a nonparametric regression setting, where a pointwise confidence interval is obtained based on the penalized series estimator. Meinshausen (2014) studies construction of one-sided confidence intervals for groups of variables under weak assumptions on the design matrix. Lockhart et al. (2014) studies significance of the input variables that enter the model along the lasso path. Lee et al. (2013) and Taylor et al. (2014) perform post-selection inference conditional on the selected model. Chatterjee et al. (2013), Liu et al. (2013), Chernozhukov et al. (2013) and Lopes (2014) study properties of the bootstrap in high-dimensions. Our work is different to the existing literature as it enables statisticians to make global inference under a nonparametric high dimensional regression setting for the first time.

1.2 Outline of the Paper

This paper is organized as follows. In the next section, we formally describe our nonparametric model, point out the issue of bias propagation when conducting inference within our model, and give two concrete examples of functions in the model family. In Section 3.2, we introduce the debiasing procedure as a remedy of bias propagation. In Section 4, we provides the theoretical results on the

statistical rate of convergence of the estimator and proposes a method for constructing confidence bands via Gaussian multiplier bootstrap.In Section 5, we apply our theoretical results to sparse additive model and sparse varying coefficient model. The numerical experiments for synthetic and real data are collected in Section 6.

1.3 Notation

Let $[n]$ denote the set $\{1, \ldots, n\}$ and let $\mathbb{1}\{\cdot\}$ denote the indicator function. For a vector $\boldsymbol{a} \in \mathbb{R}^d$, we let $\mathrm{supp}(\boldsymbol{a}) = \{j \,|\, a_j \neq 0\}$ be the support set (with an analogous definition for matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$), $\|\boldsymbol{a}\|_q$, for $q \in [1, \infty)$, the ℓ_q-norm defined as $\|\boldsymbol{a}\|_q = (\sum_{i\in[n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $\|\boldsymbol{a}\|_0 = |\mathrm{supp}(\boldsymbol{a})|$ and $\|\boldsymbol{a}\|_\infty = \max_{i\in[n]} |a_i|$. If the vector $\boldsymbol{a} \in \mathbb{R}^d$ is decomposed into groups such that $\boldsymbol{a} = (\boldsymbol{a}_{\mathcal{G}_1}, \ldots, \boldsymbol{a}_{\mathcal{G}_g})^T$, where $\mathcal{G}_1, \ldots, \mathcal{G}_g \subset [d]$ are disjoint sets, we denote $\|\boldsymbol{a}\|_{p,q}^q = \sum_{k=1}^g \|\boldsymbol{a}_{\mathcal{G}_k}\|_p^q$ and $\|\boldsymbol{a}\|_{p,\infty} = \max_{k\in[g]} \|\boldsymbol{a}_{\mathcal{G}_k}\|_p$ for any $p, q \in [1, \infty)$. We also denote the set $\{1, \ldots, j-1, j+1, \ldots, d\}$ as $\backslash j$ and the vector $\boldsymbol{a}_{\backslash j} = (a_1, \ldots, a_{j-1}, a_{j+1} \ldots, a_d)^T$. For the function $f \in L^2(\mathbb{R})$, we define the L^2 norm $\|f\|_2 = [\int f^2(x)dx]^{1/2}$ and the supremum norm $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$. For a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, we use the notation $\mathrm{vec}(\mathbf{A})$ to denote the vector in $\mathbb{R}^{n_1 n_2}$ formed by stacking the columns of $\mathbf{A}$. We denote the Frobenius norm of $\mathbf{A}$ by $||\mathbf{A}||_F^2 = \sum_{i\in[n_1], j\in[n_2]} A_{ij}^2$ and denote the operator norm as $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$. For two sequences of numbers $\{\alpha_n\}_{n=1}^\infty$ and $\{\beta_n\}_{n=1}^\infty$, we use $a_n = O(\beta_n)$ to denote that $\alpha_n \leq C\beta_n$ for some finite positive constant C, and for all n large enough. If $\alpha_n = O(\beta_n)$ and $\beta_n = O(\alpha_n)$, we use the notation $\alpha_n \asymp \beta_n$. The notation $\alpha_n = o(\beta_n)$ is used to denote that $a_n \beta_n^{-1} \xrightarrow{n\to\infty} 0$. Throughout the paper, we let c, C be two generic absolute constants, whose values may change from line to line.

2 A Family of General Nonparametric Models

We consider the problem of conducting global inference in a general nonparametric model. Specifically, we want to construct a confidence band for a nonparametric function of interest. The model we are considering here includes SpAM (**Sp**arse **A**dditive **M**odel) and sparse varying coefficient model as special cases.

First, we introduce the Hölder class of functions.

Definition 2.1 (Hölder function class)**.** The γ-th Hölder class $\mathcal{H}(\gamma, L)$ on $\mathcal{X}$ is the set of $\ell = \lfloor\gamma\rfloor$ times differentiable functions $f : \mathcal{X} \to \mathbb{R}$, where $\lfloor\gamma\rfloor$ represents the largest integer smaller than γ. The derivative $f^{(\ell)}$ satisfies

$$\left|f^{(\ell)}(x) - f^{(\ell)}(y)\right| \leq L|x - y|^{\gamma-\ell}, \quad \text{for any } x, y \in \mathcal{X}.$$

Let $\boldsymbol{X} = (X_1, \boldsymbol{X}_2^{\mathrm{T}})^{\mathrm{T}}$ be a covariate vector in $\mathcal{X}^{d+1}$, where X_1 is one dimensional random variable and $\boldsymbol{X}_2$ is d dimensional random vector. We consider the following general nonparametric model

$$Y = f(X_1)\eta(\boldsymbol{X}_2) + g(X_1, \boldsymbol{X}_2) + \varepsilon, \tag{2.1}$$

where $f \in \mathcal{H}(2, L)$ is a function of a low dimensional covariate X_1 and is treated as the component of interest, η is a known function of possibly much higher dimensional covariates $\boldsymbol{X}_2$, and g is a function of both covariates $(X_1, \boldsymbol{X}_2)$ and is treated as a nuisance. We will see that both the sparse additive model and sparse varying coefficient model are special instances of this generic nonparametric model (2.1).

We first consider the sparse additive model.

Definition 2.2 (Sparse additive model)**.** Let $\mathcal{S} \in [d]$ be of size $s = |\mathcal{S}| \ll d$. The sparse additive model can be written as

$$Y = \sum_{j \in \mathcal{S}} f_j(X_j) + \varepsilon, \tag{2.2}$$

with unknown $f_j \in \mathcal{H}(2, L)$, and ε being the error term that satisfies $\mathbb{E}[\varepsilon \,|\, \boldsymbol{X}] = 0$ almost surely.

Under the sparse additive model, we are interested in estimating and constructing a confidence band for f_1. We note that the sparse additive model can be represented as in (2.1) as follows: X_1 is the variable of interest, $(X_2, \ldots, X_d)^{\mathrm{T}}$ are represented by the nuisance variable "$\boldsymbol{X}_2$," f_1 in the sparse additive model is treated as the component of interest "f," the constant 1 as the known function "η," and $\sum_{j=2}^{d} f_j(X_j)$ as the nuisance "g."

Next, we consider the sparse varying coefficient model.

Definition 2.3 (Sparse varying coefficient model)**.** Let $\mathcal{S} \in [d]$ be of size $s = |\mathcal{S}| \ll d$. The sparse varying coefficient model can be written as

$$Y = \sum_{j \in \mathcal{S}} X_j \beta_j(Z) + \varepsilon = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}(Z) + \varepsilon, \tag{2.3}$$

with unknown $\beta_j \in \mathcal{H}(2, L)$, and ε being the error term that satisfies $\mathbb{E}[\varepsilon \,|\, \boldsymbol{X}] = 0$ almost surely. Here Z is a scalar variable that takes values in $[0, 1]$, $\boldsymbol{X} := (X_1, \ldots, X_d)^{\mathrm{T}}$ is a d dimensional vector of covariates, and $\boldsymbol{\beta}(z) = \big(\beta_1(z), \ldots, \beta_d(z)\big)^{\mathrm{T}} \in \mathcal{R}^d$ is an unknown vector-valued function of coefficients, with only $s (s \ll d)$ of its components nonzero.

Under the sparse varying coefficient model, we are interested in estimating and constructing a confidence band for β_1. We note that the sparse varying coefficient model can be represented as in (2.1) as follows: we view Z as the variable of interest "X_1," $(X_1, \ldots, X_d)^{\mathrm{T}}$ are represented by the nuisance variable "$\boldsymbol{X}_2$," β_1 is treated as the component of interest "f," X_1 as the known function "η," and $\sum_{j=2}^{d} X_j \beta_j(Z)$ as the nuisance "g."

Our goal is to obtain a consistent estimator of f in the model (2.1), and to construct a confidence band for f based on the estimator. The first goal seems easier. Had we known exactly the nuisance function g, we should be able to construct an oracle estimator of f based on minimizing the following kernel loss function

$$\widehat{f}^{\mathrm{o}}(z) := \min_{\alpha} \frac{1}{2n} \sum_{i=1}^{n} K_h(X_{i1} - z)\big(Y_i - \alpha \cdot \eta(\boldsymbol{X}_{i2}) - g(X_{i1}, \boldsymbol{X}_{i2})\big)^2. \tag{2.4}$$

A practical estimator does not have access to the unknown nuisance g and has to estimate it. We assume that a consistent initial estimator of the nuisance, which we denote as $\widehat{g}$, is obtainable. Later, we will provide a concrete estimator $\widehat{g}$ for sparse additive models and sparse varying coefficient models. With an initial estimator $\widehat{g}$, we estimate f via the following plug-in procedure

$$\widehat{f}(z) := \arg\min_{\alpha} \mathcal{L}_z(\alpha) = \frac{1}{2n} \sum_{i=1}^{n} K_h(X_{i1} - z)\big(Y_i - \alpha \cdot \eta(\boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})\big)^2. \tag{2.5}$$

We estimate the parameter of interest f by minimizing the loss function $\mathcal{L}_z(\alpha)$ in terms of α, where $\widehat{g}(\boldsymbol{X}_i)$ is the initial estimator of the nuisance.

The quality of the estimator $\widehat{f}$ depends on the rate of convergence of the initial estimator $\widehat{g}$. If the initial estimator $\widehat{g}$ is subject to bias introduced by regularization methods in high dimensional models, such bias will be propagated into $\widehat{f}$ during the plug-in estimation procedure (2.5), and therefore affecting the quality of estimator $\widehat{f}$. If the bias propagated into $\widehat{f}$ is not asymptotically ignorable, it will later affect the covering probability of confidence band constructed using $\widehat{f}$, preventing it from being an asymptotically honest one. We assume that the estimator of the nuisance $\widehat{g}(X_1, \boldsymbol{X}_2)$ in the model (2.1) has the following form

$$\widehat{g}(X_1, \boldsymbol{X}_2) = \widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{X}), \tag{2.6}$$

where $\boldsymbol{\psi}(\boldsymbol{x})^{\mathrm{T}} = \big(\boldsymbol{\psi}_1(\boldsymbol{x})^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_d(\boldsymbol{x})^{\mathrm{T}}\big)$ is a series of known basis functions, and $\widehat{\boldsymbol{\gamma}}^{\mathrm{T}} = \big(\widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\gamma}}_d^{\mathrm{T}}\big)$ is the corresponding coefficient vector. Here both $\boldsymbol{\psi}_j$ and $\widehat{\boldsymbol{\gamma}}_j$ are vectors of length m, that is, we use m basis functions when estimating each component of g. With the basis functions $\boldsymbol{\psi}_j$ $(j = 1, \ldots, d)$ fixed, we define the oracle vector $\boldsymbol{\gamma}^*$ as

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \|g(\cdot, \cdot) - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{\psi}(\cdot, \cdot)\|_{L_2}, \tag{2.7}$$

that is, we find the best approximation $\boldsymbol{\gamma}^{*\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{x})$ of the nuisance g in the L_2 norm. The approximation rate in the supremum norm $\|g - \boldsymbol{\gamma}^{*\mathrm{T}} \boldsymbol{\psi}\|_\infty$ is naturally determined by the choice of basis functions $\boldsymbol{\psi}$ and the number of bases (i.e., m) we use to approximate.

We are now ready to introduce the debiasing procedure. The debiasing procedure serves as a remedy to the previous bias propagation issue, which we will discuss more in the following section.

3 Methodology

In Section 3.1, we present the plug-in estimator $\widehat{f}$ and analyze the bias of $\widehat{f} - f$. In Section 3.2, we present the debiasing procedure. Specifically, we introduce the Immunization condition (Ning and Liu, 2014). In Section 3.3, we present the bootstrap procedures for constructing confidence bands.

3.1 Plug-in Estimator

We solve the following optimization to get plug-in estimator $\widehat{f}$:

Input Dataset $(\boldsymbol{X}, \boldsymbol{Y})$, an initial estimate $\widehat{\gamma}$, basis function $\boldsymbol{\psi}$, function η, parameter h, kernel density function K_h.
Output The plug-in estimator $\widehat{f}$ for function of interest f.

Step. Solve the follow quadratic optimization:

$$\widehat{f}(z) \leftarrow \arg\min_{\alpha} \frac{1}{2n} \sum_{i=1}^{n} K_h(X_{i1} - z)\big(Y_i - \alpha \cdot \eta(\boldsymbol{X}_{i2}) - \widehat{\gamma}^T \boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2. \tag{3.1}$$

In fact, we can have an explicit formula for $\widehat{f}$. With assumption on structure of g and $\widehat{g}$, we are now ready to study how the quality of estimation of f depends on the quality of initial estimator $\widehat{g}$. We rewrite the loss function in (2.5) as a function of $(\alpha, \boldsymbol{\gamma})$ as

$$\mathcal{L}_z(\alpha, \boldsymbol{\gamma}) := \frac{1}{2n} \sum_{i=1}^{n} K_h(X_{i1} - z)\big(Y_i - \alpha \cdot \eta(\boldsymbol{X}_{i2}) - \boldsymbol{\gamma}^{\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2. \tag{3.2}$$

Then the plug-in estimator $\widehat{f}(z) := \arg\min_\alpha \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}})$ can be viewed as the (unique) zero of $\nabla_\alpha \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}})$. As the loss function is quadratic, we have the following representation of the error

$$\widehat{f}(z) - f(z) = I(z)^{-1}\big(\boldsymbol{I}_1(z)^{\mathrm{T}}(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}) + I_2(z) + I_3(z)\big),$$

where

$$I(z) := \mathbb{E}_n\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big], \qquad \boldsymbol{I}_1(z) := \mathbb{E}_n\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(X)\big],$$
$$I_2(z) := \mathbb{E}_n\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\varepsilon\big], \quad \text{and} \quad I_3(z) := \mathbb{E}_n\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big(f(X_1) - f(z)\big)\big].$$

We notice that the initial estimation error $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ comes into estimation error $\widehat{f}(z) - f(z)$ by a factor of $I(z)^{-1}I_1(z)$. However, the estimator $\widehat{f}(z)$ is not satisfactory for the purpose of constructing a confidence band. Let's take an analytic look at error $\widehat{f}(z) - f(z)$ To construct a confidence band of f using $\widehat{f}$, we want the error introduced by $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ in $\widehat{f} - f$ to be asymptotically negligible up to the proper scaling, while here the factor $I(z)^{-1}\boldsymbol{I}_1(z)$ is obviously positive and not asymptotically negligible.

3.2 The Debiasing Procedure

In this section, we will discuss constructing confidence band based on the estimator in (3.2) In order to construct the confidence band, we will start with establish a score statistic of the formality $S(\alpha, \boldsymbol{\gamma}) := \nabla_\alpha \mathcal{L}_z(\alpha, \boldsymbol{\gamma}) - \mathbf{w}^{\mathrm{T}}\nabla_{\boldsymbol{\gamma}} \mathcal{L}_z(\alpha, \boldsymbol{\gamma})$, where $\mathbf{w} = \mathbf{w}(z)$ is a vector of the same length as $\boldsymbol{\gamma}$ and will be specified later. In particular, we will choose the vector $\mathbf{w}$ so that the following immunization condition (Ning and Liu, 2014)

$$\mathbb{E}[\nabla_{\boldsymbol{\gamma}} S(\alpha^*, \boldsymbol{\gamma}^*)] = 0, \tag{3.3}$$

holds under the model in (2.1). We obtain our debiased estimator $\widehat{\alpha} = \widehat{f}(z)$ through the solution of $S(\alpha, \widehat{\boldsymbol{\gamma}}) = 0$. In general, this condition ensures that $S(\alpha, \gamma)$ is robust to the perturbation of high dimensional nuisance parameter γ. Specifically, we have

$$S(\alpha, \widehat{\boldsymbol{\gamma}}) \approx S(\alpha, \widehat{\boldsymbol{\gamma}})$$

the difference compared to the estimator obtained by minimizing (3.2), is that the bias introduced by estimation of the nuisance component will be negligible with respect to the bandwidth of the confidence band and will not affect the limiting distribution of $\widehat{\alpha}$.

The choice of $\mathbf{w}$ is determined by the condition (3.3). From the definition of $S(\alpha, \boldsymbol{\gamma})$, we have

$$\begin{aligned}
\mathbb{E}[\nabla_{\boldsymbol{\gamma}} S(\alpha, \boldsymbol{\gamma})] &= \mathbb{E}\left[\nabla^2_{\boldsymbol{\gamma}\alpha} \mathcal{L}_z(\alpha, \boldsymbol{\gamma})\right] - \mathbf{w}^{\mathrm{T}}\mathbb{E}\left[\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \mathcal{L}_z(\alpha, \boldsymbol{\gamma})\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\boldsymbol{\psi}(\boldsymbol{X})\right] - \mathbf{w}^{\mathrm{T}}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n K_h(X_{i1} - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X})\right] - \mathbf{w}^{\mathrm{T}}\mathbb{E}\left[K_h(X_1 - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}}\right].
\end{aligned}$$

Therefore, the immunization condition (3.3) holds if $\mathbf{w}$ is set as $\mathbf{w}^*$, where

$$\mathbf{w}^* := \mathbb{E}\big[K_h(X_1 - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}}\big]^{-1}\mathbb{E}\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X})\big]. \tag{3.4}$$

We estimate $\mathbf{w}^*$ as a solution $\widetilde{\mathbf{w}} = \widetilde{\mathbf{w}}_\lambda(z)$ of the following optimization problem

$$\widetilde{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{w}\|_1 \quad \text{s.t.} \quad \Big\| n^{-1} \sum_{i=1}^{n} K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\boldsymbol{\psi}(\boldsymbol{X}_i)^{\mathrm{T}} \Big\|_{2,\infty} \leq \lambda. \tag{3.5}$$

This a convex optimization and we can solve it with some optimization software. As for λ, we just set λ of size $\mathcal{O}\big(\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}\big)$. Finally, we propose the following debiased estimator $\widetilde{f}$ of f,

$$\widetilde{f}(z) := \underset{\alpha}{\text{argzero}}\ \widetilde{S}(\alpha, \widehat{\boldsymbol{\gamma}}) = \underset{\alpha}{\text{argzero}}\ \nabla_\alpha \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}}) - \widetilde{\mathbf{w}}^{\mathrm{T}} \nabla_{\boldsymbol{\gamma}} \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}}), \tag{3.6}$$

where the estimated weight vector $\widetilde{\mathbf{w}}$ is given in (3.5), the loss function $\mathcal{L}_z$ is given in (3.2), and $\widehat{\boldsymbol{\gamma}}$ is assumed in (2.6) to be the coefficients of the basis expansion in the initial estimator $\widehat{g}$. For this debiased estimator, the bias introduced by $\gamma - \widehat{\gamma}$ can be controlled by $o(1/\sqrt{nh})$, which asymptotically negligible with respect to the bandwidth of our confidence band. Besides, we apply the "untangle and chord" procedure introduced in Ma et al. (2017). In detail, we split the whole dataset into 2 subsets randomly. We use the first subset to estimate $\widetilde{\mathbf{w}}$ and the other one to estimate $\widehat{f}$. The benefit of this procedure is that $\widetilde{\mathbf{w}}$ is independent to $\widehat{f}$, which makes $\widetilde{\mathbf{w}}$ unable to effect the normality introduced $\widehat{f}$.

The debiasing procedure consists of the following two steps:

Input Datasets $(\boldsymbol{X}', \boldsymbol{Y}'), (\boldsymbol{X}, \boldsymbol{Y})$, an initial estimate $\widehat{\gamma}$, basis function $\boldsymbol{\psi}$, function η, parameter h, kernel density function K_h.
Output The debiased estimator $\widetilde{f}$ for function of interest f.

Step 1. Estimate a row of the inverse of Hessian by solving the following optimization program:

$$\widetilde{\mathbf{w}} \leftarrow \arg\min_{\mathbf{w}} \|\mathbf{w}\|_1 \quad \text{s.t.} \quad \Big\| n^{-1} \sum_{i=1}^{n} K_h(X'_{i1} - z)\big(\eta(\boldsymbol{X}'_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}'_i)\big)\boldsymbol{\psi}(\boldsymbol{X}'_i)^{\mathrm{T}} \Big\|_{2,\infty} \leq \lambda. \tag{3.7}$$

Step 2. Obtain $\widetilde{f}$ with debiased score function:

$$\widetilde{f}(z) \leftarrow \underset{\alpha}{\text{argzero}}\ \nabla_\alpha \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}}) - \widetilde{\mathbf{w}}^{\mathrm{T}} \nabla_{\boldsymbol{\gamma}} \mathcal{L}_z(\alpha, \widehat{\boldsymbol{\gamma}}). \tag{3.8}$$

3.3 Boostrap Confidence Band

We construct the confidence band via Gaussian multiplier bootstrap with the following steps:

Input Datasets $(\boldsymbol{X}, \boldsymbol{Y})$, an initial estimate $\widehat{\gamma}$, basis function $\boldsymbol{\psi}$, function η, parameter h, kernel density function K_h, debiased estimator $\widetilde{f}$, plug-in estimator $\widehat{f}$, debiased parameter $\widetilde{w}$, quantile parameter α.

Output The confidence band $\mathcal{C}_{n,\alpha}(z)$ for function of interest $f(z)$.

Step 1. Estimate 3 statistics with debiased parameter $\widetilde{w}$:

$$\widehat{q}^2(z) \leftarrow \frac{1}{n}\sum_{i=1}^{n} K_h^2(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2, \tag{3.9}$$

$$\widehat{p}(z) \leftarrow \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(\boldsymbol{X}_{i2}), \tag{3.10}$$

$$\widehat{\sigma}^2(z) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \big(Y_i - \widehat{f}(X_{i1})\eta(\boldsymbol{X}_{i2}) - \widehat{\gamma}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2. \tag{3.11}$$

Step 2. Generate $\xi_1, ..., \xi_{n_1} \sim N(0, I_{n\times n})$. Then for a series of $z_1, ..., z_{n_2}$, we get $c_n(\alpha)$ as the $1-\alpha$ quantile of

$$\left\{ \sup_j \left| \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \xi_{ji} \frac{\widehat{\varepsilon}_i K_h(X_{i1}-z_k)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\widehat{\sigma}\widehat{q}(z_k)} \right| : k = 1, 2..., n_2 \right\},$$

where $\widehat{\epsilon}_i = Y_i - \widehat{f}(X_{i1})$. Then our confidence band is:

$$\mathcal{C}_{n,\alpha}(z) := \widetilde{f}(z) \pm \widehat{c}_n(\alpha) n^{-1/2} \widehat{\sigma}\widehat{q}(z)/\widehat{p}(z). \tag{3.12}$$

4 Theoretical Results

This section establishes the rate of convergence for the proposed estimator in Section 3.2. Based on the debiased estimator in (3.6), a confidence band for f is constructed via a Gaussian multiplier bootstrap.

4.1 Assumptions

We start with stating the required assumptions. Let $p_{\boldsymbol{X}}(\boldsymbol{x})$ denote the joint density of $\boldsymbol{X} = (X_1, \boldsymbol{X}_2)$ and let $p_{X_1}(x_1)$ denote the marginal density of X_1. The b and B below refer to constants such that $0 < b \le B < \infty$, while their values are subject to change in different assumptions.

(A1) (Density function) The support $\mathcal{X}$ is compact and the density function $p_{\boldsymbol{X}}(\boldsymbol{x})$ is continuous on $\mathcal{X}^d$. There exist fixed constants b, B such that $b \le p_{X_1}(x_1) \le B$ for all $x_1 \in \mathcal{X}$. The function $\eta(\cdot)$ depends only finite (say, s) components of $\boldsymbol{x}_2$ (say, $(x_{21}, \ldots, x_{2s})$), and the joint density of $(X_1, X_{21}, \ldots, X_{2s})$ is bounded away from 0: $p_{X_1,\boldsymbol{X}_{2,1:s}}(x_1, \boldsymbol{x}_{2,1:s}) \ge b$.

(A2) (Kernel function) The kernel $K(u)$ is a continuous density function with bounded variation, satisfying

$$\int_{\mathcal{X}} K(u)du = 1 \quad \text{and} \quad \int_{\mathcal{X}} uK(u)du = 0.$$

(A3) (Basis function) The basis functions $\boldsymbol{\psi}$ are uniformly bounded on $\mathcal{X}$

$$\max_{j\in[d],k\in[m]} \|\boldsymbol{\psi}_{jk}\|_\infty \le B.$$

(A4) (Noise term) The error term ε satisfies $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$, $\mathbb{E}[\varepsilon^2 \mid \boldsymbol{X}] = \sigma^2$, and is a subgaussian random variable such that $\mathbb{E}[\exp(\lambda\varepsilon)] \leq \exp(\lambda^2\sigma_\varepsilon^2/2)$ for any λ.

(A5) (Nonparametric model) The nonparametric function $\mathbb{E}[Y \mid \boldsymbol{X}]$ has the form in (2.1), with the function $\eta(\cdot)$ satisfying

$$\|\eta\|_\infty \leq B \quad \text{and} \quad \|\eta\|_{L_2} \geq b.$$

(A6) (Rate of approximation and estimation) The nuisance component $g(\cdot)$ can be well approximated with the basis functions $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_d)$ as the rate r_a, that is

$$\|g - \boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}\|_\infty \leq Cr_a. \tag{4.1}$$

The initial estimator $\widehat{\boldsymbol{\gamma}}$ satisfies

$$\mathbb{P}\big(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_{2,1} > Cr_e\big) \leq 1/n. \tag{4.2}$$

(A7) (Debiased term) To guarantee the existence of $\mathbf{w}^*$ in (3.4), we assume the minimal eigenvalue of the Hessian matrix

$$\Sigma_z^* = \begin{bmatrix} \mathbb{E}K_h(X_1 - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}} & \mathbb{E}K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}} \\ \mathbb{E}K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X}) & \mathbb{E}K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2 \end{bmatrix}$$

is larger than a constant b for all z.

(A8) (Nonparametric sparsity) $\|\mathbf{w}^*\|_1 \leq L$.

Assumption **(A1)** is on the density function of covariates. Under sparse additive model $h \equiv 1$, hence h involves 0 component of $\boldsymbol{x}_2$, and the assumption $p_{X_1,\boldsymbol{X}_{2,1:s}}(x_1, \boldsymbol{x}_{2,1:s}) \geq b$ automatically holds when $b = 1$ is chosen. Under sparse varying coefficient model $h(x_1) = x_1$, hence h involves only 1 variable of $\boldsymbol{x}_2$, and the assumption $p_{X_1,\boldsymbol{X}_{2,1:s}}(x_1, \boldsymbol{x}_{2,1:s}) \geq b$ is equivalent to the joint density of (Z, X_1) is bounded away from 0, i.e., $p_{Z,X_1}(z, x_1) \leq b$. Under both cases, **(A1)** seems reasonable.

Assumption **(A2)** is standard in the literature on local linear regression (Fan, 1993). Assumption **(A3)** automatically holds if we use normalized B-spline basis as $\boldsymbol{\psi}$, and is hence reasonable. Assumption **(A4)** is standard for the noise variable in SpAM (Meier et al., 2009; Huang et al., 2010; Koltchinskii et al., 2010; Zhang et al., 2011; Raskutti et al., 2012; Kato, 2012).

Assumption **(A5)** is on the boundedness of function η. It is reasonable as we want the multiplicative factor in our model, function η, to be strong enough for f to be estimable.

Assumption **(A6)** is on the rate of basis approximation and initial estimation. For sparse additive model, we can choose $r_a = sm^{-2}$, and $r_e = n^{-1/2}m\sqrt{\log(md)} + m^{-3/2}$. For sparse varying coefficient model, we can choose $r_a = sm^{-2}$, and $r_e = n^{-1/2}m\log(md)$.

Assumption **(A7)** are conditions to guarantee the existence of $\mathbf{w}^*$. This is the same as the Assumption 4.2 in Ning and Liu (2014). Recall that $\mathbf{w}^* := \mathbb{E}\big[K_h(X_1 - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}}\big]^{-1}\mathbb{E}\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X})\big]$. In terms of Schur components, if Σ_z^* is nonsingular, $\mathbb{E}\big[K_h(X_1 - z)\boldsymbol{\psi}(\boldsymbol{X})\boldsymbol{\psi}(\boldsymbol{X})^{\mathrm{T}}\big]^{-1}$ is nonsingular too, which guarantees the existence of $\mathbf{w}^*$. Besides, we also have

$$\begin{aligned}\mathbb{E}K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2 - \mathbf{w}^{*T}\mathbb{E}K_h(X_1 - z)\eta(\boldsymbol{X}_2)\boldsymbol{\psi}(\boldsymbol{X}) &= \begin{bmatrix} -\mathbf{w}^{*T} & 1 \end{bmatrix}\Sigma_z \begin{bmatrix} -\mathbf{w}^* \\ 1 \end{bmatrix} \\ &\geq b\sqrt{\|\mathbf{w}^*\|_2 + 1} \geq b.\end{aligned}$$

We will mainly use this inequality in our proofs instead of **(A7)**.

Assumption **(A8)**

4.2 Estimation Consistency

We present the rate of convergence of the plug-in estimator.

Theorem 4.1 (Convergence rate of $\widehat{f}$ to f)**.** Under Assumption 4.1, there exists a constant C such that if $(nh)^{-1}\log n = \mathcal{O}(1)$ and $h = o(1)$ $m \to \infty$ as $n \to \infty$ and we have that the plug-in estimator $\widehat{f}$ defined in (2.5) satisfies

$$\sup_{z\in\mathcal{X}} \left|\widehat{f}(z) - f^*(z)\right| \le C\Big(r_a + \sqrt{n^{-1}h\log n} + h^2 + (nh)^{-1/2}\log n + \sqrt{m}r_e\Big), \tag{4.3}$$

with probability $1 - c/n$ for a constant $c > 0$.

4.3 Asymptotically Honest Confidence Band via Gaussian Multiplier Bootstrap

Recently, Chernozhukov et al. (2014a) and Chernozhukov et al. (2014b) developed a novel framework to analyze the Gaussian multiplier bootstrap for approximating the distribution of empirical processes and apply it to kernel density estimation. We extend their approach and consider the Gaussian multiplier bootstrap to estimate the distribution of the supremum of a studentized process

$$Z_n(z) := \frac{\sqrt{n}\big(\widetilde{f}(z) - \mathbb{E}_{f^*}[\widetilde{f}(z)]\big)}{\sigma(z)}, \tag{4.4}$$

where $\sigma(z)$ is a normalizing factor to be determined.

The idea here is straightforward. As long as the bias term $\mathbb{E}_{f^*}[\widetilde{f}(z)] - f^*(z)$ is asymptotically negligible up to the normalizing factor $\sigma(z)$, we can construct a confidence band of the form $\widetilde{f}(z) \pm c_n(\alpha)n^{-1/2}\sigma(z)$, where $c_n(\alpha)$ is the $(1-\alpha)$-quantile of supremum of the process $Z_n(z)$ (or its approximated version). Besides, Ma et al. (2017) also proposed a novel method for debiased estimator. Thus, we separate the dataset into subsets and use one to estimate $\widetilde{\mathbf{w}}$ and another one to solve our debiased estimator. The benefit of this method is that $\widetilde{\mathbf{w}}$ can be regarded as a constant in our proof and it won't damage our normality since it's estimated by another set. By setting $\sigma(z)$ as $\sigma q(z)/\widehat{p}(z)$ and removing all bias terms, we obtain an empirical process

$$\mathbb{G}_n(z) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\varepsilon_i K_h(X_{i1} - z)\big(\eta(X_{i2}) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\sigma q(z)}, \tag{4.5}$$

where $q^2(z) := \mathbb{E}\big[K_h(X_1 - z)^2\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X})\big)^2\big]$, and $\widehat{p}(z) = n^{-1}\sum_{i=1}^{n} K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)$.

In order to estimate the distribution of supremum of $\mathbb{G}_n$, in particular its $(1-\alpha)$-quantile, we consider the following calculable Gaussian multiplier process

$$\widehat{\mathbb{G}}_n(z) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \xi_i \cdot \frac{\widehat{\varepsilon}_i K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\widehat{\sigma}\widehat{q}(z)}, \tag{4.6}$$

where $\{\xi_i\}_{i=1}^n$ are independent standard Gaussian random variables which are independent of $(\boldsymbol{X}, \boldsymbol{\varepsilon})$, and $\widehat{\varepsilon}_i := Y_i - \widehat{f}(X_{i1})\eta(\boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})$ is the estimator of the noise term ε_i. The terms σ and $q(z)$ in denominator of $\mathbb{G}_n(z)$ are also replaced by their estimators $\widehat{\sigma}^2 := n^{-1}\sum_{i=1}^{n}\big(Y_i - \widehat{f}(X_{i1})\eta(\boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})\big)^2$ and $\widehat{q}(z)^2 := n^{-1}\sum_{i=1}^{n} K_h(X_{i1} - z)^2\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2$.

Finally, let $\widehat{c}_n(\alpha)$ be the conditional $(1-\alpha)$-quantile of $\sup_{z\in\mathcal{X}}\big|\widehat{\mathbb{G}}_n(z)\big|$, on fixed $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$. Based on the debiased estimator $\widetilde{f}(z)$, we construct the confidence band at level $(1-\alpha)$ as $\mathcal{C}_{n,\alpha} = \{\mathcal{C}_{n,\alpha}(z)|z\in\mathcal{X}\}$ where

$$\mathcal{C}_{n,\alpha}(z) := \widetilde{f}(z) \pm \widehat{c}_n(\alpha) n^{-1/2} \widehat{\sigma}\widehat{q}(z)/\widehat{p}(z). \tag{4.7}$$

We can show that this confidence band is asymptotically honest with polynomial convergence rate.

Here we present several conditions about parameters r_e, r_a, m, h, d to guarantee our confidence band is honest.

Theorem 4.2 (Asymptotically honest confidence band via Gaussian multiplier bootstrap)**.** Define the function class $\mathcal{F} = \{f \in \mathcal{F}(2, L) | \|f\|_\infty < M\}$. Under Assumption 4.1, assume $m \asymp n^{\delta_1}, h \asymp n^{-\delta_2}$, where $0 < \delta_1 < 1, \frac{1}{5} < \delta_2 < \frac{1}{4}$. If we also have $2\delta_1 + \delta_2 < 1, \log(dn)\cdot n^{\delta_1 - 1 + \delta_2} \to 0, r_a \cdot n^{(1-\delta_2)/2} \to 0, r_e \cdot n^{\delta_1} \to 0, \log(dn)\cdot r_e \to 0$, then there exist constants $C, c > 0$ such that we have the coverage probability of $f^*(z)$

$$\inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha - Cn^{-c}. \tag{4.8}$$

In particular, the confidence band $\mathcal{C}_{n,\alpha}$ is asymptotically honest, that is,

$$\liminf_{n\to\infty} \inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha.$$

Remark 4.3. The confidence band constructed by the multiplier bootstrap neither requires ε to be symmetric nor relies on any conditions on the existence of limiting distribution for the supremum of the studentized empirical process in (4.4). We will give a proof of Theorem 4.2 in Section 7.

5 Convergence rates for Specific Model Classes

In this section, we give some theoretical results about convergence of $\widehat{f}$ and confidence bands $\mathcal{C}_{n,\alpha}$ when applied to particular subclasses of the nonparametric model in (2.1). In particular, we provide debiasing estimators for sparse additive models and sparse varying coefficient models and corollaries for the convergences of their estimators and confidence bands.

5.1 Sparse Additive Models

We first present the debiased estimator for the sparse additive model (2.2). We approximate f_j, $j = 2, \ldots, d$ by a linear combination of the normalized B-spline basis functions $\{\phi_1, \ldots, \phi_m\}$. Given basis functions, we denote f_{jm} as the projection (in terms of L_2 norm) of f_j onto the space spanned by the basis. We have

$$f_{jm} := \mathop{\arg\min}_{f\in\mathcal{S}_m} \|f - f_j\|_2 = \sum_{k=1}^m \gamma_{jk}^* \psi_{jk},$$

where $\mathcal{S}_{jm} = \mathrm{Span}(\psi_{j1}, \ldots, \psi_{jm})$ and $\psi_{jk} := \phi_k(x) - \mathbb{E}_n[\phi_k(X_j)]$ are centered basis functions. With this, we approximate the nuisance $g(\mathbf{X}) = \sum_{j=2}^d f_j(X_j)$ as

$$g(\mathbf{X}) \approx \sum_{j=2}^d f_{jm}(X_j) = \sum_{j=2}^d \sum_{k=1}^m \gamma_{jk}^* \psi_{jk}(X_j) = (\boldsymbol{\gamma}_2^{*\mathrm{T}}, \ldots, \boldsymbol{\gamma}_d^{*\mathrm{T}})\big(\boldsymbol{\psi}_2(X_2)^\mathrm{T}, \ldots, \boldsymbol{\psi}_d(X_d)^\mathrm{T}\big)^\mathrm{T},$$

where $\boldsymbol{\gamma}_j^* = \big(\gamma_{j1}^*, \ldots, \gamma_{jm}^*\big)^{\mathrm{T}}$ and $\boldsymbol{\psi}_j(x_j) = \big(\psi_{j1}(x_j), \ldots, \psi_{jm}(x_j)\big)^{\mathrm{T}}$. For functions in the Hölder class $\mathcal{H}(2, L)$, we have that $\|f_j - f_{jm}\|_\infty = \mathcal{O}(m^{-2})$ (see equation (20) in Zhou et al. (1998)). Furthermore, due to the sparsity assumption in our sparse additive model, we have that

$$\Big\| \sum_{j=2}^{d} (f_j - f_{jm}) \Big\|_\infty = \mathcal{O}(sm^{-2}). \tag{5.1}$$

We estimate $\boldsymbol{\gamma}^*$ using the group lasso estimator

$$\widehat{\boldsymbol{\gamma}} := \arg\min_{\boldsymbol{\gamma}} \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\gamma}\|_2^2 + \lambda \sum_{j=1}^{d} \|\boldsymbol{\gamma}_j\|_2, \quad \text{where } \boldsymbol{\Psi}^{(n\times md)} = [\boldsymbol{\psi}(\boldsymbol{X}_1), \ldots, \boldsymbol{\psi}(\boldsymbol{X}_n)]^{\mathrm{T}}, \tag{5.2}$$

which was proposed in Huang et al. (2010). When the number of nonzero components s is fixed and the tuning parameter $\lambda \asymp \sqrt{n \log(md)}$, we have that (see Theorem 1 in Huang et al., 2010)

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 = \mathcal{O}_{\mathbb{P}}\Big(\frac{m^2 \log(md)}{n} + \frac{1}{m^3} + \frac{m^2\lambda^2}{n^2}\Big)$$

and

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_{2,1} := \sum_{i=2}^{d} \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2 = \mathcal{O}_{\mathbb{P}}\Big(m\sqrt{\frac{\log(md)}{n}} + m^{-3/2}\Big). \tag{5.3}$$

Applying Theorem 4.2 to the SpAM model, we have:

Corollary 5.1 (Confidence band under sparse additive model)**.** For sparse additive model, we choose $r_a = sm^{-2}, r_e = n^{-1/2}m\sqrt{\log(md)} + m^{-3/2}$. Under Assumption 4.1, if $m \asymp n^{\delta_1}, 3/16 < \delta_1 < 1/4$, $h \asymp n^{-\delta_2}, 1/5 < \delta_2 < 1/4$ and as $n \to \infty$, $s \cdot n^{1/2-\delta_2/2-2\delta_1} \to 0, n^{-1+4\delta_1} \cdot \log(nd) \to 0$, there exists constants $C, c > 0$ such that we have the coverage probability of $f^*(z)$

$$\inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha - Cn^{-c}. \tag{5.4}$$

In particular, the confidence band $\mathcal{C}_{n,\alpha}$ is asymptotically honest, that is,

$$\liminf_{n\to\infty} \inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha.$$

Here we compare our estimator and confidence bands to those in Lu et al. (2019). There are three major differences between them: Firstly, their paper proposed a one-step plug-in estimator while our estimator is two-step. Secondly, their assumption for design matrix is specialized for B-splines, while we only assume our Hessian matrices' minimal eigenvalues are larger than a constant in Assumption 4.1 and our theoretic results can be applied to other first-step estimators for nuisance. Thirdly, about the biasing process, we apply the "untangle and chord" procedure introduced in Ma et al. (2017). In detail, we split the dataset into two subsets and use one subset to estimate $\widetilde{\mathbf{w}}$ and the other one to estimate $\widehat{f}$. While in Lu et al. (2019), they estimate $\widetilde{\mathbf{w}}$ ($\widehat{\theta}$ in their paper) and $\widehat{f}$ with the same dataset.

5.2 Sparse Varying Coefficient Model

In sparse varying coefficient model (2.3), again we can approximate β_j by B-spline basis. Given the normalized basis functions $\{\phi_1, \ldots, \phi_m\}$, we denote β_{jm} as the projection (in terms of L_2 norm) of β_j onto the space spanned by the basis

$$\beta_{jm} := \arg\min_{\beta \in \mathcal{S}_m} \|\beta - \beta_j\|_2 = \sum_{k=1}^{m} \gamma_{jk}^* \phi_k, \quad \text{where } \mathcal{S}_m = \text{Span}(\phi_1, \ldots, \phi_m).$$

Hence the nuisance $\sum_{j=2}^{d} \beta_j(Z) X_j$ can be approximated by

$$\begin{aligned}\sum_{j=2}^{d} \beta_j(Z) X_j \asymp \sum_{j=2}^{d} \beta_{jm}(Z) X_j &= \sum_{j=2}^{d} \sum_{k=1}^{m} \gamma_{jk}^* \phi_k(Z) X_j \\ &= (\boldsymbol{\gamma}_2^{*\mathrm{T}}, \ldots, \boldsymbol{\gamma}_d^{*\mathrm{T}}) \big(\boldsymbol{\psi}_2(Z, X_2)^{\mathrm{T}}, \ldots, \boldsymbol{\psi}_d(Z, X_d)^{\mathrm{T}}\big)^{\mathrm{T}},\end{aligned}$$

where $\boldsymbol{\gamma}_j^* = \big(\gamma_{j1}^*, \ldots, \gamma_{jm}^*\big)^{\mathrm{T}}$ and $\boldsymbol{\psi}_j(z, x_j) = \big(x_j \phi_1(z), \ldots, x_j \phi_m(z)\big)^{\mathrm{T}}$. For the same reason as above, we have the approximation rate (in $\|\cdot\|_\infty$) using B-spline basis on Hölder class $\mathcal{H}(2, L)$

$$\|\beta_j - \beta_{jm}\|_\infty = \mathcal{O}(m^{-2}), \quad \forall f_j \in \mathcal{H}(2, L),$$

and similarly as in (5.1)

$$\Big\| \sum_{j=1}^{d} \big(x_j \beta_j(z) - x_j \beta_{jm}(z)\big) \Big\|_\infty = \mathcal{O}(s m^{-2}). \tag{5.5}$$

Now let us consider the estimation rate of $\boldsymbol{\gamma}^*$. We assume the number of nonzero components s is fixed. According to Theorem 3 in Lian (2012), the group lasso estimator $\widehat{\boldsymbol{\gamma}}$

$$\widehat{\boldsymbol{\gamma}} := \arg\min_{\boldsymbol{\gamma}} \|\boldsymbol{Y} - \boldsymbol{\Psi}^{\mathrm{T}} \boldsymbol{\gamma}\|_2^2 + \lambda \sum_{j=1}^{d} \|\boldsymbol{\gamma}_j\|_1, \quad \text{where } \boldsymbol{\Psi}^{(n \times md)} = [\boldsymbol{\psi}(Z_1, \boldsymbol{X}_1), \ldots, \boldsymbol{\psi}(Z_n, \boldsymbol{X}_n)]^{\mathrm{T}} \tag{5.6}$$

with tuning parameter $\lambda / \sqrt{n \log d \vee n} \to \infty$ and $\lambda m^2 / n \to 0$, satisfies

$$\sum_{i=1}^{d} \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2^2 = \mathcal{O}_{\mathbb{P}}\left(m^2 \lambda^2 / n^2\right).$$

This implies we have the estimation rate in terms of $\|\cdot\|_{2,1}$ norm

$$\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_{2,1} := \sum_{i=2}^{d} \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2 = \mathcal{O}_{\mathbb{P}}\left(m\lambda / n\right). \tag{5.7}$$

Applying Theorem 4.2 to the SVCM model, we have:

Corollary 5.2 (Confidence band under sparse varying model)**.** For sparse varying model, we choose $r_a = s m^{-2}, r_e = n^{-1/2} m \log(md)$. Under Assumption 4.1, if $m \asymp n^{\delta_1}, 3/16 < \delta_1 < 1/4$,

$h \asymp n^{-\delta_2}, 1/5 < \delta_2 < 1/4$ and as $n \to \infty$, $s \cdot n^{1/2-\delta_2/2-2\delta_1} \to 0, n^{-1/2+2\delta_1} \cdot \log(nd) \to 0$, there exist constants $C, c > 0$ such that we have the coverage probability of $f^*(z)$

$$\inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha - Cn^{-c}. \tag{5.8}$$

In particular, the confidence band $\mathcal{C}_{n,\alpha}$ is asymptotically honest, that is,

$$\liminf_{n\to\infty} \inf_{f^*(z)\in\mathcal{F}} \mathbb{P}\big(f^*(z) \in \mathcal{C}_{n,\alpha}(z), \forall z \in \mathcal{X}\big) \geq 1 - \alpha.$$

6 Numerical Experiments

In this section, we study the finite sample properties of confidence bands for sparse additive model and sparse varying coefficient model.

Example 6.1. We generate data from the following sparse additive model

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i = \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the additive components are designed as follows

$$f_1(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3),$$
$$f_2(t) = 5t, \quad f_3(t) = 3(2t-1)^2, \quad f_4(t) = 4\sin(2\pi t)/(2 - \sin(2\pi t)),$$

and $f_5(t) = \ldots = f_d(t) = 0$. The noise $\varepsilon_i \sim N(0, 1.5^2)$ for $i = 1, \ldots, n$. The number of nonzero functions is thus $s = 4$. This generating model is the same as Example 1 of Huang et al. (2010). Let $W_1, \ldots, W_d$ and U follow iid Uniform$[0, 1]$ and

$$X_j = \frac{W_j + tU}{1 + t} \quad \text{for } j = 1, \ldots, d.$$

The covariates $X_{1j}, X_{2j}, \ldots, X_{2j}$ are iid copies of X_j. We set the dimensions of covariates $d = 600$, $t = 0.3$, and consider three different sample sizes $n = 100, 300, 500$. In the initial group lasso estimator (2.6), we use the cubic B-splines with six evenly distributed knots and $m = 9$. The tuning parameter λ and bandwidth h are chosen by cross validation according to the BIC criterion defined as

$$\mathrm{BIC}(\lambda, h) = \log(\mathrm{RSS}_{(\lambda,h)}) + \mathrm{df}_{(\lambda,h)} \cdot \frac{\log n}{n},$$

where RSS is the residual sum of squares for a given (λ, h), and the degrees of freedom is defined as $\mathrm{df} = \widehat{s} \cdot m$ with $\widehat{s}$ being the number of nonzero estimated components selected for given (λ, h). We aim at constructing confidence band for $f_1^*(t) = f_1(t) - \mathbb{E}[f_1(X_1)]$. To test the coverage probability of confidence bands for inactive covariates, we also construct confidence band for $f_5(t) \equiv 0$. The confidence band in (4.7) is constructed at the significance level 95% and the quantile estimator $\widehat{c}_n(\alpha)$ is computed by bootstrap with 1,000 repetitions. To measure the coverage probability of the confidence band, we compute empirical probability that the confidence band covers the true function on the first 100 data points based on 500 repetitions. The numerical results are reported in Figure 1 and Table 1.

Example 6.2. We generate data from the following sparse varying coefficient model

$$Y_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}(Z_i) + \varepsilon_i = a_1 \cdot X_{i1}\beta_1(Z_i) + \sum_{j=1}^{d} X_{ij}\beta_j(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where we set

$$a_1 \in \{0, 1\}, \quad \beta_1(z) = 3 + \sin(2\pi z), \quad \beta_2(z) = \exp(z + 1), \quad \beta_3(z) = 4(2z - 1)^2,$$

and $\beta_4(z) = \ldots = \beta_d(z) = 0$. Here two values of $a_1 \in \{0, 1\}$ correspond to two scenarios that the true function is zero and nonzero. The noise $\varepsilon_i \sim N(0, 1.5^2)$ for $i = 1, \ldots, n$. The number of nonzero functions is thus $s = 3$. This generating model is adapted from Example 1 of Lian (2012) by adding a_1. The index variable Z is sampled uniformly on $[0, 1]$, and the covariates $X_{i1} = 1$ fixed and $X_{i2}, \ldots, X_{id}$ are independently and generated from $U(-1, 1)$, $i = 1, \ldots, n$. We set the dimensions of covariates $d = 1000$ and consider three different sample sizes $n = 100, 300, 500$. In the initial group lasso estimator (2.6), we use the cubic B-splines with six evenly distributed knots and $m = 9$. Again we tune λ and h through cross validation via minimizing the BIC criterion. We aim to construct the confidence band for $a_1 f_1(t)$. The confidence band in (4.7) is constructed at the significance level 95% and the quantile estimator $\widehat{c}_n(\alpha)$ is computed by bootstrap with 1,000 repetitions. To measure the coverage probability of the confidence band, we compute empirical probability that the confidence band covers the true function on the first 100 data points based on 500 repetitions. The numerical results are reported in Figure 2 and Table 1.

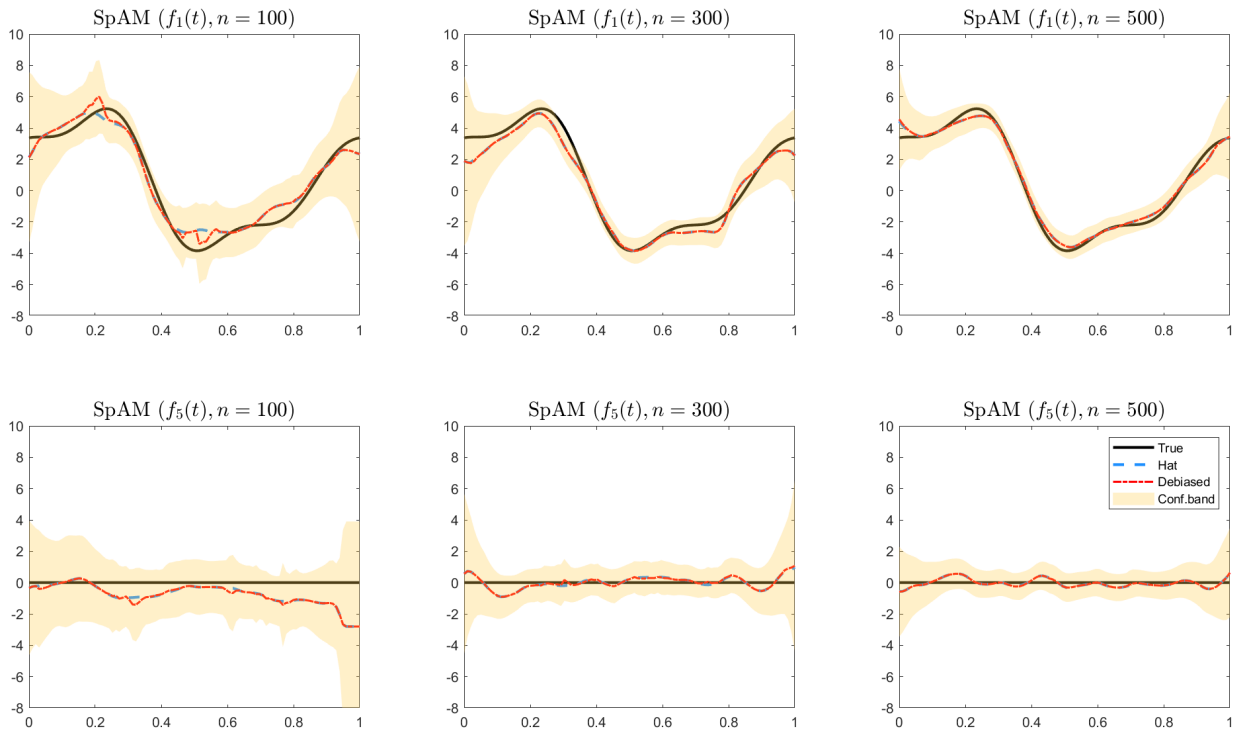

Figure 1: Debiased estimators for the $d = 600$ dimensional SpAM model $Y = \sum_{j=1}^{4} f_j(X_j) + \varepsilon$, for $n = 100, 300, 500$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant bands at significant level 95% cover $f_1(t)$ on the first row and $f_5(t) = 0$ on the second row.

		Scheme 1: f_1			Scheme 2: f_5		
	Sample size:	100	300	500	100	300	500
SpAM	Cover probability	0.68	0.93	0.94	0.82	0.92	0.96
		Scheme 1: $a_1 = 1$			Scheme 2: $a_1 = 0$		
	Sample size:	100	300	500	100	300	500
SpVC	Cover probability	0.85	0.88	0.96	0.89	0.92	0.97

Table 1: Coverage probabilities for confidence bands at significant level 95% for SpAM model $Y = \sum_{j=1}^{4} f_j(X_j) + \varepsilon$ and sparse varying coefficient model $Y = \sum_{j=1}^{3} X_j \beta_j(Z) + \varepsilon$ with dimension $d = 1000$, sample size $n = 100, 300, 500$, and $\varepsilon \sim N(0, 1.5^2)$.

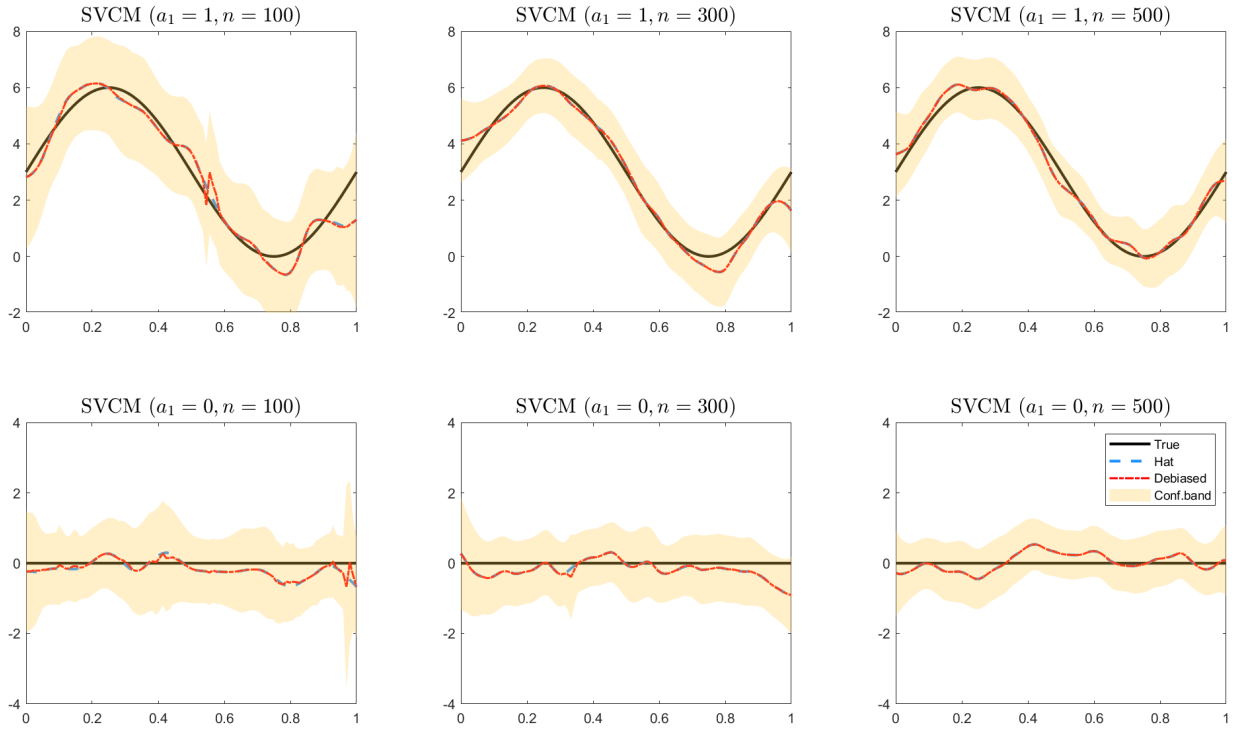

Figure 2: Debiased estimators for the $d = 600$ dimensional SVCM model $Y = \sum_{j=1}^{3} X_j \beta_j(Z) + \varepsilon$, for $n = 100, 300, 500$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant bands at significant level 95% cover $f_1^* = a_1 f_1$ for $a_1 \in \{0, 1\}$, respectively.

7 Proofs of Main Theoretical Results

We collects the proofs of main theorems in this section. The proofs of Theorem 4.1 and Theorem 4.2 is presented in Section 7.1 and Section 7.2 respectively.

7.1 Proof of the Statistical Rate of the Plug-in Estimator

This section outlines the proof of Theorem 4.1 on the statistical estimation rate of the plug-in estimator in (2.5).

The estimation error for the estimator $\widehat{f}$ comes from four sources: (1) noise ε, (2) approximation error by basis expansion, (3) estimation error introduced by kernel methods, and (4) error introduced by initial estimator $\widehat{\boldsymbol{\gamma}}$. Before presenting the main proof, we list several technical lemmas whose proofs are deferred to Appendix A in the supplementary material. The following lemma provides the rate for basis approximation. We denote $\mathbf{W}_z := \operatorname{diag}\big(K_h(X_{11}-z),\ldots,K_h(X_{n1}-z)\big) \in \mathbb{R}^{n\times n}$ as the kernel matrix, and $\boldsymbol{\eta} := \big(\eta(\boldsymbol{X}_{12}),\ldots,\eta(\boldsymbol{X}_{n2})\big)^{\mathrm{T}}$ below.

Lemma 7.1. Let $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_n)^{\mathrm{T}}$, where $\delta_i = g(X_{i1},\boldsymbol{X}_{i2}) - \boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)$ for $i=1,\ldots,n$. Under Assumptions 4.1 and $(nh)^{-1}\log n = \mathcal{O}(1)$, there exists a constant $C>0$, such that the following inequality holds with probability at least $1-1/n$,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\delta}| \le Cr_a, \tag{7.1}$$

where r_a is the approximation rate of basis $\boldsymbol{\psi}$: $\|g-\boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}\|_\infty \le Cr_a$.

Our next lemma bound the estimation error introduced by kernel methods. We can see that the smoothing parameter(i.e., bandwidth) h play a role in the estimation.

Lemma 7.2. Let $\boldsymbol{\xi}_z = (\xi_1,\ldots,\xi_n)^{\mathrm{T}}$, where $\xi_i(z) = \big(f(X_{i1})-f(z)\big)\eta(\boldsymbol{X}_{i2})$ for $i=1,\ldots,n$. Under Assumptions 4.1 and $(nh)^{-1}\log n = \mathcal{O}(1)$, there exists a constant $C>0$, such that the following inequality holds with probability at least $1-1/n$,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\xi}_z| \le Ch^2 + C\sqrt{n^{-1}h\log n}. \tag{7.2}$$

The following lemma quantifies the statistical rate from the noise ε.

Lemma 7.3. Let $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^{\mathrm{T}}$. Under Assumption 4.1 and $(nh)^{-1}\log n = \mathcal{O}(1)$, there exists a constant $C>0$, such that the following inequality holds with probability at least $1-4/n$,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\varepsilon}| \le C(nh)^{-1/2}\log n. \tag{7.3}$$

The last lemma quantifies the statistical rate introduced by initial estimation error of $\widehat{\boldsymbol{\gamma}}$.

Lemma 7.4. Let $\boldsymbol{\zeta} = (\zeta_1(z),\ldots,\zeta_n(z))^{\mathrm{T}}$, where $\zeta_i = \big(\boldsymbol{\gamma}^* - \widehat{\boldsymbol{\gamma}}\big)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)$. Under Assumption 4.1 and $(nh)^{-1}\log n = \mathcal{O}(1)$, there exists a constant $C>0$, such that the following inequality holds with probability at least $1-1/n$,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\zeta}| \le C\sqrt{m}r_e, \tag{7.4}$$

where r_e is the estimation rate of $\boldsymbol{\gamma}^*$ in ℓ_1 norm: $\mathbb{P}\big(\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_{2,1} > Cr_e\big) \le 1/n$.

We are now ready to present the main proof of Theorem 4.1.

Proof of Theorem 4.1. Recall that $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^{\mathrm{T}}$, $\boldsymbol{\xi}_z = (\xi_1(z), \ldots, \xi_n(z))^{\mathrm{T}}$, $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)^{\mathrm{T}}$, where $\delta_i = g(X_{i1}, \boldsymbol{X}_{i2}) - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)$, $\xi_i(z) = \big(f(X_{i1}) - f(z)\big)\eta(\boldsymbol{X}_{i2})$, $\zeta_i = \big(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\big)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)$, for $i = 1, \ldots, n$. We denote $\boldsymbol{\eta}_z := \boldsymbol{\varepsilon} + \boldsymbol{\delta} + \boldsymbol{\xi}_z + \boldsymbol{\zeta}$.

$$\begin{aligned}\widehat{f}(z) - f(z) &= \frac{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\big(Y_i - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})\big)}{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})^2} - f(z) \\ &= \frac{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\big(Y_i - f(z)\eta(\boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})\big)}{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})^2}.\end{aligned}$$

Plugging in (2.1) and (2.6), we have

$$\begin{aligned}\widehat{f}(z) - f(z) &= \frac{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\Big(\big(f(X_{i1}) - f(z)\big)\eta(\boldsymbol{X}_{i2}) + g(X_{i1}, \boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2}) + \varepsilon_i\Big)}{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})^2}, \\ &= \frac{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})\Big[\big(f(X_{i1}) - f(z)\big)\eta(\boldsymbol{X}_{i2}) + g(X_{i1}, \boldsymbol{X}_{i2}) - \widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i) + \varepsilon_i\Big]}{\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})^2} \\ &= \frac{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z(\boldsymbol{\xi}_z + \boldsymbol{\delta} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon})}{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}} = \frac{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}_z}{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}}.\end{aligned}$$

Using Lemma 7.1, Lemma 7.2, Lemma 7.3, and Lemma 7.4, there exist a constant C' such that with probability $1 - 8/n$,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}_z| \le C' r_n, \tag{7.5}$$

where $r_n := r_a + \sqrt{n^{-1}h\log n} + h^2 + (nh)^{-1/2}\log n + \sqrt{m}r_e$.

By McDiarmid's inequality, there exists a constant C, such that

$$\mathbb{P}\bigg(\Big|n^{-1}\sum_{i=1}^n K_h(X_{i1} - z)\eta(\boldsymbol{X}_{i2})^2 - \mathbb{E}\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big]\Big| \ge C\sqrt{\log n}/(\sqrt{n}h)\bigg) \le \frac{1}{n}.$$

Also by **(A1)** and **(A5)** in Assumption 4.1, we have

$$\begin{aligned}\mathbb{E}[\eta(\boldsymbol{X}_2)^2|X_1 = x] &= \int \eta(\boldsymbol{X}_{2,1:s})^2 \frac{p_{X_1,\boldsymbol{X}_{2,1:s}}(x, \boldsymbol{x}_{2,1:s})}{p_{X_1}(x)} d\boldsymbol{x}_{2,1:s} \ge B^{-1}b\|h\|_{L_2}^2 \ge B^{-1}b^3. \\ \mathbb{E}\big[K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big] &= \mathbb{E}\big[K_h(X_1 - z)\mathbb{E}[\eta(\boldsymbol{X}_2)^2|X_1]\big] \ge B^{-1}b^3 \cdot \mathbb{E}[K_h(X_1 - z)],\end{aligned}$$

where $\mathbb{E}[K_h(X_1 - z)] = \int_{\mathcal{X}} K(t)p_{X_1}(z + th)dt \ge b/2$. Hence with probability at least $1 - 1/n$, we have

$$\frac{1}{n}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta} \ge B^{-1}b^4/4, \quad \forall z \in \mathcal{X}. \tag{7.6}$$

To sum up, on the event where both (7.5) and (7.6) hold (with probability at least $1 - 9/n$), we have

$$\sup_{z\in\mathcal{X}}\Big|\widehat{f}(z) - f^*(z)\Big| = \sup_{z\in\mathcal{X}} \frac{n^{-1}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}_z|}{n^{-1}\boldsymbol{\eta}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\eta}} \le 4BC'/b^4 \cdot r_n = Cr_n.$$

□

7.2 Proof of Covering Properties of the Bootstrap Confidence Bands

In this section, we prove Theorem 4.2 on the coverage probabilities for the Gaussian Multiplier bootstrap confidence bands $\mathcal{C}_{n,\alpha}$ in (4.7).

In our proof, we will apply the the "untangle and chord" procedure introduced in Ma et al. (2017) to get our $\widetilde{\boldsymbol{w}}$, Thus, we use $\{(\boldsymbol{X}_i', Y_i')\}, i = 1, ..., n$ to denote the subdataset for $\mathbf{w}$ and $\{(\boldsymbol{X}_i, Y_i)\}$ to denote the subdataset for our debiased estimator. Recall that we have two optimization problems for getting $\widetilde{f}$:

$$\begin{aligned} \widetilde{\mathbf{w}} &= \arg\min \|\mathbf{w}\|_1 \\ \text{s.t.} \quad &\|\nabla^2_{\alpha\gamma} L_z - \mathbf{w}^T \nabla^2_{\gamma\gamma} L_z\|_{2,\infty} \le \lambda, \end{aligned} \tag{7.7}$$

where we use the subdataset $\{(\boldsymbol{X}_i', Y_i')\}$ and $\lambda_1 = C\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}$. Then,

$$\begin{aligned} \widetilde{f} &= \text{argzero} S_z(\alpha, \widehat{\gamma}) \\ \text{s.t.} \quad &S_z(\alpha, \widehat{\gamma}) = \nabla_\alpha L_z(\alpha, \widehat{\gamma}) - \widetilde{\mathbf{w}} \nabla_\gamma L_z(\alpha, \widehat{\gamma}), \end{aligned} \tag{7.8}$$

where we use the subdataset $\{(\boldsymbol{X}_i, Y_i)\}$. Now we list all the lemmas for the main proof. The first lemma guarantees the existence of solution to Optimization 7.7.

Lemma 7.5. Under Assumptions 4.1 and conditions in Theorem 4.2, with probability $1 - \frac{c}{n}$, there exists a solution $\widetilde{\mathbf{w}}$ for the optimization problem 7.7, and $|\widetilde{\mathbf{w}}|_1 \le L$.

For convenience, we also use $\widetilde{\alpha}, \alpha^*$ to denote $\widetilde{f}(z), f^*(z)$. Then the following lemma gives the difference between $\widetilde{\alpha}$ and α^*.

Lemma 7.6. According to the optimization problems 7.7 and 7.8, under the condition that $(nh)^{-1}\log n = o(1)$, we have

$$\widetilde{\alpha} - \alpha^* = \frac{(\nabla^2_{\alpha\gamma} L_z - \widetilde{\mathbf{w}}^T \nabla^2_{\gamma\gamma} L_z)(\gamma^* - \widehat{\gamma}) + \widetilde{\mathbf{w}} \nabla_r L_z(\alpha^*, \gamma^*) - \nabla_\alpha L_z}{\nabla^2_{\alpha\alpha} L_z - \widetilde{\mathbf{w}} \nabla^2_{\alpha\gamma} L_z}.$$

Then we use I_1, I_2 to denote

$$\frac{(\nabla^2_{\alpha\gamma} L_z - \widetilde{\mathbf{w}}^T \nabla^2_{\gamma\gamma} L_z)(\gamma^* - \widehat{\gamma})}{\nabla^2_{\alpha\alpha} L_z - \widetilde{\mathbf{w}} \nabla^2_{\alpha\gamma} L_z}, \quad \frac{\widetilde{\mathbf{w}} \nabla_r L_z(\alpha^*, \gamma^*) - \nabla_\alpha L_z(\alpha^*, \gamma^*)}{\nabla^2_{\alpha\alpha} L_z - \widetilde{\mathbf{w}} \nabla^2_{\alpha\gamma} L_z}.$$

Thus, $\widetilde{\alpha} - \alpha^* = I_1 + I_2$. The following lemma upper bounds I_1.

Lemma 7.7. Under Assumption 4.1, there exists a constant C such that

$$\mathbb{P}\left(I_1 \le C r_e \cdot \frac{m + \sqrt{m\log(dnh^{-1})}}{\sqrt{nh}} \right) \ge 1 - \frac{7}{n}.$$

We can rewrite I_2 as

$$\begin{aligned} I_2 &= \frac{\frac{1}{n}\sum_{i=1}^n K_h(X_{i1} - z)(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T \boldsymbol{\psi}(\boldsymbol{X}_i))(Y_i - \alpha^* \eta(\boldsymbol{X}_{i2}) - \gamma^{*T} \boldsymbol{\psi}(\boldsymbol{X}_i))}{\frac{1}{n}\sum_{i=1}^n K_h(X_{i1} - z)(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T \boldsymbol{\psi}(\boldsymbol{X}_i))\eta(\boldsymbol{X}_{i2})} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)[(f^*(X_{i1}) - f(z))\eta(\boldsymbol{X}_{i2}) + g(\boldsymbol{X}_i) - \gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)]}{\widehat{p}(z)} + \frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)\varepsilon_i}{\widehat{p}(z)}, \end{aligned}$$

where $t(\boldsymbol{X}_i, z) \coloneqq K_h(X_{i1}-z)(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i))$ and $\widehat{p}(z) \coloneqq n^{-1}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)$. Now we use I_{21}, I_{22} to denote,

$$\frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)[(f^*(X_{i1})-f(z))\eta(\boldsymbol{X}_{i2})+g(\boldsymbol{X}_i)-\gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)]}{\widehat{p}(z)}, \quad \frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)\varepsilon_i}{\widehat{p}(z)}.$$

The following lemma gives an upper bound of I_{21}.

Lemma 7.8. Under Assumption 4.1, there exist constants C and c, such that

$$\mathbb{P}\left(I_{21} \le C(r_a + h^2 + \sqrt{n^{-1}h\log n})\right) \ge 1 - \frac{c}{n}.$$

In fact $I_{21}+I_1$ is the bias term in our debiased estimator. We will proof it's asymptotically negligible with respect to our bandwidth of confidence band later. As for I_{22}, we will established a sequence of processes to construct confidence band for its numerator. Recall that

$$\mathbb{G}_n(z) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i K_h(X_{i1}-z)\big(\eta(X_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\sigma q(z)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\varepsilon_i t(\boldsymbol{X}_i, z)}{\sigma q(z)},$$
$$\widehat{\mathbb{G}}_n(z) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \cdot \frac{\widehat{\varepsilon}_i K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\widehat{\sigma}\widehat{q}(z)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\xi_i\widehat{\varepsilon}_i t(\boldsymbol{X}_i, z)}{\widehat{\sigma}\widehat{q}(z)},$$

where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i^2$ is the estimator of $\sigma^2 = \mathrm{Var}(\varepsilon_1)$, and $\widehat{q}(z)^2 = n^{-1}\sum_{i=1}^n K_h(X_{i1}-z)^2\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2$ is the estimator of $q^2(z) = \mathbb{E}\big[K_h(X_1-z)^2\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X})\big)^2\big]$. The following lemma gives a bound for the difference between $\mathbb{G}_n(z)$'s and $\widehat{\mathbb{G}}_n(z)$'s denominators. Let

$$r_n \coloneqq r_a + \sqrt{n^{-1}h\log n} + h^2 + (nh)^{-1/2}\log n + r_e\sqrt{m}.$$

Lemma 7.9. Let the estimator for $\mathrm{Var}(\varepsilon) = \sigma^2$ be $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{\varepsilon}_i^2$. Under Assumption 4.1, there exist constants C and c such that $\mathbb{P}(|\widehat{\sigma}^2 - \sigma^2| \ge Cr_n) \le 6/n$ and

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}}\left|\frac{\widehat{\sigma}\widehat{q}(z)}{\sigma q(z)} - 1\right| \le C\left(r_n + \sqrt{\log n/(nh)}\right)\right) \ge 1 - \frac{c}{n}.$$

Then, to approximate $\mathbb{G}_n(z)$ by $\widehat{\mathbb{G}}_n(z)$, we need two intermediate processes

$$\mathbb{G}_n^{(1)}(z) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \Bigg[\frac{\varepsilon_i I(|\varepsilon_i| \le b_n)}{\sigma}\cdot\frac{K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{q(z)}$$
$$-\mathbb{E}\Big(\frac{\varepsilon_i I(|\varepsilon_i| \le b_n)}{\sigma}\cdot\frac{K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{q(z)}\Big)\Bigg], \tag{7.9}$$

$$\mathbb{G}_n^{(2)}(z) \coloneqq \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\xi_i\varepsilon_i I(|\varepsilon \le b_n)\cdot K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\sigma q(z)}, \tag{7.10}$$

where $b_n \coloneqq C'\sqrt{\log n}$ such that subgaussian random variables ε_i's satisfy $\mathbb{P}(\max_{i\in[n]}\varepsilon_i \le b_n) \ge 1-n^{-1}$. The next two lemmas show how well $\mathbb{G}_n^{(1)}(z), \mathbb{G}_n^{(2)}(z)$ approximates $\mathbb{G}_n(z)$ and $\widehat{\mathbb{G}}_n(z)$. Assume $\boldsymbol{X}_1^n = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}, \varepsilon_1^n = \{\varepsilon_1, ..., \varepsilon_n\}$ and $\boldsymbol{J}_n = (\boldsymbol{X}_1^n, \varepsilon_1^n)$. We denote

$$W_n \coloneqq \sup_{z\in\mathcal{X}}\mathbb{G}_n(z), \quad W_n^{(1)} \coloneqq \sup_{z\in\mathcal{X}}\mathbb{G}_n^{(1)}(z), \quad W_n^{(2)}(\boldsymbol{j}_n) \coloneqq \sup_{z\in\mathcal{X}}\mathbb{G}_n^{(2)}(z) \text{ and } \widehat{W}_n(\boldsymbol{j}_n) \coloneqq \sup_{z\in\mathcal{X}}\widehat{\mathbb{G}}_n(z),$$

for $\boldsymbol{j}_n \in \Omega \coloneqq \mathbb{R}^n \times (\mathcal{X}^d)^n$. Then we have the following lemmas:

Lemma 7.10. Under conditions in Theorem 4.2, there exists constants c_1, c_2, c such that

$$\mathbb{P}(|W_n - W_n^{(1)}| > c_1 n^{-c}) \leq c_2 n^{-c}. \tag{7.11}$$

Lemma 7.11. Under conditions in Theorem 4.2, there exists constants c_1, c_2, c and a set $S_n \subset \Omega$ satisfying $\mathbb{P}(\boldsymbol{J}_n \in S_n) \geq 1 - \frac{3}{n}$ For any fixed $\boldsymbol{j}_n \in S_n$, we have

$$\mathbb{P}(|W_n^{(1)} - W_n^{(2)}(\boldsymbol{j}_n)| > c_1 n^{-c}) \leq c_2 n^{-c}. \tag{7.12}$$

Lemma 7.12. Under conditions in Theorem 4.2, there exists constants c_1, c_2, c and a set $S_n \subset \Omega$ satisfying $\mathbb{P}(\boldsymbol{J}_n \in S_n) \geq 1 - \frac{3}{n}$. For any fixed $\boldsymbol{j}_n \in S_n$, we have

$$\mathbb{P}(|W_n^{(2)}(\boldsymbol{j}_n) - \widehat{W}_n(\boldsymbol{j}_n)| > c_1 n^{-c}) \leq c_2 n^{-c}. \tag{7.13}$$

Now we can give the main proof concerning the convergence our confidence band.

Proof of Theorem 4.2. Recall our confidence band is

$$\mathcal{C}_{n,\alpha}(z) := \widetilde{f}(z) \pm \widehat{c}_n(\alpha) n^{-1/2} \widehat{\sigma}\widehat{q}(z)/\widehat{p}(z).$$

Firstly, we want to show the biased term in $\widetilde{f}$ is asymptotically with respect to the bandwidth of $\mathcal{C}_{n,\alpha}(z)$. Applying Lemma B.8, Lemma B.9 and $\widehat{p}(z) > b/4$, we could see the width of the band satisfying

$$2\widehat{c}_n(\alpha) n^{-1/2} \widehat{\sigma}\widehat{q}(z)/\widehat{p}(z) \geq C \frac{1}{\sqrt{nh}}. \tag{7.14}$$

Applying Lemma 7.7 and Lemma 7.8, the biased term in $\widetilde{f}$ is $I_1 + I_{21}$, and under conditions in Theorem 4.2, with probability $1 - \frac{c}{n}$,

$$I_1 + I_{21} \leq C\left(r_e \cdot \frac{m + \sqrt{m \log(dnh^{-1})}}{\sqrt{nh}} + r_a + h^2 + \sqrt{n^{-1} h \log n}\right) = o\left(\frac{1}{\sqrt{nh}}\right). \tag{7.15}$$

Combine (7.14) and (7.15), the bias in $\widetilde{f}$ is asymptotically negligible. Since $\widetilde{f} - f^* = I_1 + I_{21} + I_{22}$, and

$$I_{22} = \frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)\varepsilon_i}{\widehat{p}(z)},$$

we will construct the confidence band for the numerator of I_{22}, $\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)\varepsilon_i$. After normalizing the variance of the numerator, we get

$$\mathbb{G}_n(z) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\varepsilon_i t(\boldsymbol{X}_i, z)}{\sigma q(z)}.$$

We hope to approximate $\mathbb{G}_n(z)$ with $\widehat{\mathbb{G}}_n(z)$,

$$\widehat{\mathbb{G}}_n(z) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\xi_i \widehat{\varepsilon}_i t(\boldsymbol{X}_i, z)}{\widehat{\sigma}\widehat{q}(z)}.$$

We use the Corollary 3.1 in Chernozhukov et al. (2014a) to establish this approximation. Corollary 3.1 of Chernozhukov et al. (2014a) provides sufficient conditions for the confidence band to be asymptotically honest. Specifically, we verify the following high-level conditions (Let C_1 be some positive constant):

H1 (Gaussian approximation). There exists a sequence of random variables W_n^0 such that $W_n^0 \stackrel{d}{=} \sup_{z\in\mathcal{X}} \mathbb{G}_n^0(z)$, where $\mathbb{G}_n^0$ is a Gaussian process in $\ell^\infty(\mathcal{X})$. Furthermore, $\mathbb{E}\big[\sup_{z\in\mathcal{X}} \mathbb{G}_n^0(z)\big] \le c_1\sqrt{\log n}$ and

$$\sup_{f^*\in\mathcal{F}} \mathbb{P}\big(|W_n - W_n^0| > \epsilon_{1n}\big) \le \delta_{1n},$$

for some ϵ_{1n} and δ_{1n} bounded from above by $c_1 n^{-c}$.

H2 (Anti-concentration). For any $\epsilon > 0$, the anti-concentration inequality

$$\sup_{x\in\mathbb{R}} \mathbb{P}\left(\Big|\sup_{z\in\mathcal{X}} |\mathbb{G}_n(z)| - x\Big| \le \epsilon\right) \le C_1\epsilon\sqrt{\log n}$$

holds.

H3 (Estimation error of $\widehat{c}_n(\alpha)$). Let $c_n(\alpha)$ be the $(1-\alpha)$-quantile of $W_n^0 = \sup_{z\in\mathcal{X}} G_n(z)$ and $\widehat{c}_n(\alpha)$ be the $(1-\alpha)$-quantile of $\widehat{W}_n = \sup_{z\in\mathcal{X}} \widehat{G}_n(z)$. For some τ_n, ε_{2n} and δ_{2n} bounded from above by $c_1 n^{-c}$, we have

$$\sup_{f^*\in\mathcal{F}} \mathbb{P}\Big(\widehat{c}_n(\alpha) < c_n(\alpha+\tau_n) - \varepsilon_{2n}\Big) \le \delta_{2n},$$

and

$$\sup_{f^*\in\mathcal{F}} \mathbb{P}\Big(\widehat{c}_n(\alpha) > c_n(\alpha-\tau_n) + \varepsilon_{2n}\Big) \le \delta_{2n}.$$

H4 (Estimation error of $\widehat{\sigma}_n(\cdot)$). For some ε_{3n} and δ_{3n} bounded from above by $c_1 n^{-c}$, we have

$$\sup_{f^*\in\mathcal{F}} \mathbb{P}\left(\sup_{z\in\mathcal{X}} \left|\frac{\widehat{\sigma}\widehat{q}(z)}{\sigma q(z)} - 1\right| > \varepsilon_{3n}\right) \le \delta_{3n}.$$

First, we verify the condition **H1**. Under conditions in Theorem 4.2, Lemma 7.10 gives us the the rate of difference between $W_n^{(1)}$ and W_n

$$\mathbb{P}\big(|W_n - W_n^{(1)}| > c_1 n^{-c}\big) \le c_2 n^{-c}. \tag{7.16}$$

Then we will construct a good Gaussian approximation to the intermediate random variable $W_n^{(1)}$. As for $\mathbb{G}_n^{(1)}$, the function class of interest is:

$$\mathcal{S}_h := \left\{ S_z(\varepsilon, \boldsymbol{X}) := \frac{\varepsilon I(|\varepsilon| \le b_n)}{\sigma} \cdot \frac{K_h(X_1 - z)\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X})\big)}{q(z)} \Big| z \in \mathcal{X} \right\},$$

where $b_n = C\sqrt{\log n}$. We are going to prove $\mathbb{S}_h$ is a VC-type class (Definition C.4). As for the upper bound,

$$\|S_z(\varepsilon, \boldsymbol{X})\|_\infty \le \frac{C\sqrt{\log n}(L+1)B}{\sqrt{h}\sigma b} = C\sqrt{\frac{\log n}{h}}.$$

Therefore, one parameter for VC-type class is $b_{VC} = C\sqrt{\frac{\log n}{h}}$. According to Lemma C.2, for any $\tau > 0$ and all finite measure $\mathbb{Q}$,

$$N(\mathcal{S}_h, L_2(\mathbb{Q}), b_{VC}\tau) \le \left(\frac{2\|K\|_{tv}A}{\tau}\right)^4,$$

where A is an absolute constant. Thus the parameters of $\mathcal{S}_h$ are

$$\begin{cases} v_{VC} = 4 = \mathcal{O}(1), \\ a_{VC} = 2\|K\|_{tv} A = \mathcal{O}(1), \\ b_{VC} = \mathcal{O}(\sqrt{\log n \cdot h^{-1}}), \\ \sigma_{VC}^2 = \mathbb{E} S_z^2(\varepsilon, \boldsymbol{X}) = \mathcal{O}(1), \end{cases}$$

and

$$K_h := A v_{VC}(\log n \vee \log(a_{VC} b_{VC}/\sigma_{VC}) = \mathcal{O}(\log n).$$

Applying Lemma C.5, there exists a tight Gaussian random element $B = \{B(g) : g \in S_h\}$ such that

$$\mathbb{E}\big[B(g_1)B(g_2)\big] = \mathbb{E}\big[g_1(\boldsymbol{X}_1, \varepsilon_1) g_2(\boldsymbol{X}_1, \varepsilon_1)\big] - \mathbb{E}\big[g_1(\boldsymbol{X}_1, \varepsilon_1)\big]\mathbb{E}\big[g_2(\boldsymbol{X}_1, \varepsilon_1)\big], \tag{7.17}$$

for all $g_1, g_2 \in S_h$. Let $\mathbb{G}_n^0(z) = B(S_z)$ and $W^0 := \sup_z B(S_z)$. We have for any $\gamma \in (0, 1)$,

$$\mathbb{P}\left(|W_n^{(1)} - W^0| > C \cdot \left(\frac{\log^2 n}{\gamma^{1/2} h n^{1/2}} + \frac{\log^{5/4} n}{\gamma^{1/2} n^{1/4} h^{3/4}} + \frac{\log n}{\gamma^{1/3} n^{1/6} h^{2/3}} \right) \right) \leq A' \Big(\gamma + \frac{\log n}{n} \Big),$$

where A' is an absolute constant. Under conditions in Theorem 4.2, since $h \asymp n^{-\delta_2}, \delta_2 < 1/4$, there exists c such that $\frac{1}{n^{1/6} h^{2/3}} \leq n^{-c}$. Hence, with $\gamma = n^{-3c/2}$, there exist constants c_1, c_2 such that

$$\mathbb{P}\Big(|W_n^{(1)} - W^0| > c_1 n^{-c} \Big) \leq c_2 n^{-c}. \tag{7.18}$$

Then combining (7.16) and (7.18), we have

$$\mathbb{P}\Big(|W_n - W^0| > c_1 n^{-c} \Big) \leq c_2 n^{-c}.$$

Now we are going to verify $\mathbb{E}[\sup_{z \in \mathcal{X}} \mathbb{G}_n^0(z)] = \mathbb{E}[\sup_{z \in \mathcal{X}} B(S_z)] \leq c_1 \sqrt{\log n}$. With (7.17), we know the covering number for $\mathbb{G}_n^0(z)$ is the same as $\mathbb{G}_n^{(1)}(z)$. Therefore, applying Dudley's inequality (Dudley, 2010), we have

$$\begin{aligned} \mathbb{E}[\sup_{z \in \mathcal{X}} \mathbb{G}_n^0(z)] &\leq C \int_0^\infty \sqrt{\log N(\mathcal{S}_h, L^2(\mathbb{Q}), \tau)}\, d\tau \\ &\leq C \int_0^C \sqrt{2 \log C \frac{\log n}{h} + 4 \log \frac{1}{\tau}}\, d\tau = C \sqrt{\log n}. \end{aligned}$$

The condition **H2** follows from **H1** and the anti-concentration inequality in Corollary 2.1 of Chernozhukov et al. (2014a).

Next, we verify the condition **H3**. Recall that $\widehat{c}_n(\alpha)$ is the $(1-\alpha)$-quantile of $\widehat{W}_n$. From the approximation result in (7.12), (7.13) and (7.18), there exists a subset S_n, such that $\mathbb{P}(\boldsymbol{J}_n \in S_n) \geq 1 - \frac{3}{n}$. For any $\boldsymbol{j}_n \in S_n$, we have

$$\mathbb{P}\big(|\widehat{W}_n(\boldsymbol{j}_n) - W_n^0| > c_1 n^{-c} + \epsilon_{1n}' \big) \leq c_2 n^{-c} + \delta_{1n}',$$

and hence with another two constants c_1' and c_2', we have,

$$\mathbb{P}\big(|\widehat{W}_n(\boldsymbol{j}_n) - W_n^0| > c_1' n^{-c}\big) \leq c_2' n^{-c}.$$

Therefore, we can bound the probability

$$\begin{aligned}\mathbb{P}\big(W_n^0 \leq \widehat{c}_n(\alpha) + c_1' n^{-c}\big) &\geq \mathbb{P}\big(\widehat{W}_n(\boldsymbol{j}_n) \leq \widehat{c}_n(\alpha)\big) - \mathbb{P}\big(|\widehat{W}_n(\boldsymbol{j}_n) - W_n^0| > c_1' n^{-c}\big) \\ &\geq 1 - \alpha - c_2' n^{-c}, \end{aligned} \tag{7.19}$$

which implies that the estimated quintile has the following lower bound

$$\widehat{c}_n(\alpha) \geq c_n(\alpha + c_2' n^{-c}) - c_1' n^{-c}. \tag{7.20}$$

Similarly, we also have $\widehat{c}_n(\alpha) \leq c_n(\alpha - c_2' n^{-c}) + c_1' n^{-c}$. By setting $\tau = c_2' n^{-c}$, $\epsilon_{2n} = c_1' n^{-c}$, and $\delta_{2n} = 0$, condition **H3** holds.

Finally, using Lemma 7.9, the condition **H4** is satisfied for $\epsilon_{3n} = C\big(r_n + \sqrt{\log n/(nh)}\big) = o(n^{-c})$ and $\delta_{3n} = 8/n$. Since the high-level condition **H1** − **H4** are verified, the result follows from Corollary 3.1 in Chernozhukov et al. (2014a).

□

References

Avalos, M., Grandvalet, Y. and Ambroise, C. (2007). Parsimonious additive models. *Computational statistics & data analysis* **51** 2851–2870.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *Ann. Statist.* **46** 3643–3675.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014a). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

Belloni, A., Chernozhukov, V. and Kato, K. (2013a). Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv preprint arXiv:1312.7186* .

Belloni, A., Chernozhukov, V. and Kato, K. (2014b). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika* asu056.

Belloni, A., Chernozhukov, V. and Wei, Y. (2013b). Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969* .

Bernstein, S. (1964). On a modification of chebyshev's inequality and on the error in laplace formula. *Collected Works, Izd-vo'Nauka', Moscow (in Russian)* **4** 71–80.

Bertin, K., Lecué, G. et al. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics* **2** 1224–1241.

Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30** 207–216.

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique* **334** 495–500.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: minimax rates and adaptivity. *Ann. Statist.* **45** 615–646.

Cai, T. T. and Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 391–419.

Chatterjee, A., Lahiri, S. et al. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics* **41** 1232–1259.

Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* **41** 2786–2819.

Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics* **42** 1787–1818.

Chernozhukov, V., Chetverikov, D., Kato, K. et al. (2014b). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics* **42** 1564–1597.

Comminges, L., Dalalyan, A. S. et al. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics* **40** 2667–2696.

Dalalyan, A., Ingster, Y. and Tsybakov, A. B. (2014). Statistical inference in compound functional models. *Probability Theory and Related Fields* **158** 513–532.

Dudley, R. M. (2010). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes [mr0220340]. In *Selected works of R. M. Dudley*. Sel. Works Probab. Stat., Springer, New York, 125–165.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics* 196–216.

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

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics* 1491–1518.

Farrell, M. H. (2013). Robust inference on average treatment effects with possibly more covariates than observations. *Available at SSRN 2324292* .

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association* **76** 817–823.

Giné, E. and Nickl, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probability Theory and Related Fields* **143** 569–596.

Gregory, K., Mammen, E. and Wahl, M. (2016). Optimal estimation of sparse high-dimensional additive models. *Preprint* .

Guo, Z. and Zhang, C.-H. (2019a). Extreme nonlinear correlation for multiple random variables and stochastic processes with applications to additive models. *arXiv preprint arXiv:1904.12897* .

Guo, Z. and Zhang, C.-H. (2019b). Local inference in additive models with decorrelated local linear estimator. *arXiv preprint arXiv:1907.12732* .

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)* 757–796.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, vol. 43. CRC Press.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics* **38** 2282.

Javanmard, A. and Montanari, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274* .

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

Kato, K. (2012). Two-step estimation of high dimensional additive models. *arXiv preprint arXiv:1207.5313* .

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, vol. 2033. Springer Science & Business Media.

Koltchinskii, V., Yuan, M. et al. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics* **38** 3660–3695.

Kozbur, D. (2013). Inference in additively separable models with a high dimensional component. Tech. rep., Working Paper. Booth School of Business, University of Chicago, Chicago, IL.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics* 28–63.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2013). Exact post-selection inference with the lasso. *arXiv preprint arXiv:1311.6238* .

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica* **22** 1563.

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

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636.

Liu, H., Yu, B. et al. (2013). Asymptotic properties of lasso+ mls and lasso+ ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* **7** 3124–3169.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *Annals of statistics* **42** 413.

Lopes, M. (2014). A residual bootstrap for high-dimensional regression with near low-rank designs. In *Advances in Neural Information Processing Systems*.

Lou, Y., Bien, J., Caruana, R. and Gehrke, J. (2014). Sparse partially linear additive models. *arXiv preprint arXiv:1407.4729* .

Lu, J., Kolar, M. and Liu, H. (2019). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association* 1–40.

Ma, C., Lu, J. and Liu, H. (2017). Inter-subject analysis: Inferring sparse interactions with dense intra-graphs. *arXiv preprint arXiv:1709.07036* .

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

Meinshausen, N. (2014). Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417–473.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**.

Ning, Y. and Liu, H. (2014). Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity. *arXiv preprint arXiv:1412.2295* .

Petersen, A., Witten, D. and Simon, N. (2014). Fused lasso additive model. *arXiv preprint arXiv:1409.5391* .

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

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

Rosasco, L., Villa, S., Mosci, S., Santoro, M. and Verri, A. (2013). Nonparametric sparsity and regularization. *The Journal of Machine Learning Research* **14** 1665–1714.

Sardy, S. and Tseng, P. (2004). Amlet, ramlet, and gamlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics* **13** 283–309.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 55–80.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics* 689–705.

Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *arXiv preprint arXiv:1401.3889* .

Vaart, A. v. d. and Wellner, J. A. (1997). Weak convergence and empirical processes with applications to statistics. *Journal of the Royal Statistical Society-Series A Statistics in Society* **160** 596–608.

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

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

Wahl, M. (2014). Variable selection in high-dimensional additive models based on norms of projections. *arXiv preprint arXiv:1406.0052* .

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics* **37** 2178.

Xu, M., Chen, M. and Lafferty, J. (2014). Faithful variable screening for high-dimensional convex regression. *arXiv preprint arXiv:1411.1805* .

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

Yao, Y. and Zhang, C.-H. (2020). Adaptive estimation in high-dimensional additive models with multi-resolution group lasso. *arXiv preprint arXiv:2011.06765* .

Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.* **44** 2564–2593.

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

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

A Auxiliary Lemmas for Estimation Results

Throughout the appendix we use C to denote an absolute constant, whose value may change from line to line. The following technical result will be useful throughout the proofs.

Lemma A.1. With probability $1 - \frac{1}{n}$, there exist a constant C such that

$$\sup_z \frac{1}{n} \sum_{i=1}^{n} K_h(X_{i1} - z) \leq C.$$

Proof. Consider the process and the function class of interest:

$$U_n(z) = \frac{1}{n} \sum_{i=1}^{n} K_h(X_{i1} - z) - \mathbb{E}K_h(X_{i1} - z)$$
$$\mathcal{S}_h = \{S_z(x_1, \boldsymbol{x}_2) := K_h(x_1 - z) | z \in \mathcal{X}\}.$$

According to Lemma C.2,

$$N\left(\mathcal{S}_h, L^2\left(\mathbb{P}_n\right), \epsilon\right) \leq \left(\frac{2\|K\|_{\mathrm{TV}} A}{h\epsilon}\right)^4,$$

where A is an absolute constant. Then we can upper bound the variance of the process as:

$$\sigma_p^2 := \sup_z \mathbb{E}K_h(X_{i1} - z)^2 = \mathcal{O}(h^{-1}).$$

Therefore, according to Lemma C.1,

$$\mathbb{E}\left[\sup_{z \in \mathcal{X}} U_n(z)\right] \leq C n^{-1/2} \sigma_p \sqrt{\log \frac{2\|K\|_{TV}}{h\sigma_p}} \leq C\sqrt{\frac{\log h^{-1}}{nh}}. \tag{A.1}$$

Since $\sup_z |K_h(x_1 - z)| \leq U := Ch^{-1}$, by Lemma C.3, we have

$$\mathbb{P}\left(\sup_z U_n(z) \geq \mathbb{E}\left[\sup_z U_n(z)\right] + t\sqrt{2\left(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\left[\sup_z U_n(z)\right]\right) + \frac{2Ut^2}{3}}\right) \leq \exp\left(-nt^2\right). \tag{A.2}$$

Combine (A.1) and (A.2), set $t = \sqrt{\frac{\log n}{n}}$, then, with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned}\sup_z U_n(z) &\leq \mathbb{E}\left[\sup_z U_n(z)\right] + \sqrt{2\sigma_{\mathrm{P}}^2 + 4U\mathbb{E}\left[\sup_z U_n(z)\right]} \cdot \sqrt{\frac{\log n}{n}} + \frac{2U}{3} \cdot \frac{\log n}{n} \\ &\leq C\sqrt{\frac{\log n}{nh}} \leq C,\end{aligned}$$

where we used the inequality $\sqrt{xy} \leq 2^{-1}(x + y)$ and the assumption that $(nh)^{-1} \log n = \mathcal{O}(1)$. □

The estimation error for the plug-in estimator $\widehat{f}$ comes from four sources: (1) noise, (2) approximation error by finite B-spline bases, (3) bias introduced by the local constant estimator, and (4) estimation error from the initial estimator $\widehat{\boldsymbol{\gamma}}$.

The following lemma provides the rate of the B-spline approximation error. It shows how the number of B-spline basis functions m influences the final estimation rate.

Proof of Lemma 7.1. By Assumption 4.1, for any $i \in [n]$, $|\eta(\boldsymbol{X}_{i2})| \le \|h\|_\infty$ and $|\delta_i| = |g(X_{i1}, \boldsymbol{X}_{i2}) - \boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)| \le \|g - \boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}\|_\infty \le r_a$ are uniformly bounded random variables. First,

$$\sup_{z\in\mathcal{X}} \frac{1}{n}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\delta}| \le \sup_{z\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})|\delta_i|.$$

We denote the empirical process and the function class of interest as

$$U_{n,\boldsymbol{\delta}}(z) = \frac{1}{n}\sum_{i=1}^n \eta(\boldsymbol{X}_{i2})K_h(X_{i1}-z)|\delta_i| - \mathbb{E}\big[\eta(\boldsymbol{X}_{12})K_h(X_{11}-z)|\delta_1|\big],$$
$$\mathcal{S}_{h,\boldsymbol{\delta}} = \left\{ s_z(x_1,\boldsymbol{x}_2) = \eta(\boldsymbol{x}_2)h^{-1}K\big(h^{-1}(x_1-z)\big)|\delta(x_1,\boldsymbol{x}_2)| \big| z\in\mathcal{X}\right\},$$

where $\delta(\boldsymbol{x}) = g(\boldsymbol{x}) - \boldsymbol{\gamma}^{*\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})$ satisfies $|\delta(\boldsymbol{x})| \le r_a$ for any $\boldsymbol{x} = (x_1, \boldsymbol{x}_2)$. For simplicity, the notation $U_n(z)$ and $\mathcal{S}_h$ is used instead of $U_{n,\boldsymbol{\delta}}(z)$ and $\mathcal{S}_{h,\boldsymbol{\delta}}$, when it causes no confusion. Let $\mathcal{F}_h = \{h^{-1}K\big(h^{-1}(\cdot - z)\big)|z\in\mathcal{X}\}$ and let $\|K\|_{\mathrm{TV}}$ be the total variation of $K(\cdot)$. From Lemma C.2, for all probability measures Q on $\mathcal{X}$,

$$N(\mathcal{F}_h, L^2(Q), \epsilon) \le \left(\frac{2\|K\|_{\mathrm{TV}}A}{h\epsilon}\right)^4,$$

where A is an absolute constant and $0<\epsilon<1$. Let $\widetilde{\mathcal{F}}_h$ be an $\|h\|_\infty^{-1}r_a^{-1}\epsilon$-cover of $\mathcal{F}_h$ with respect to $\widehat{\mathbb{P}}_{X_1} = n^{-1}\sum_{i=1}^n \delta_{X_{i1}}$. We then construct an ϵ-covering for $\mathcal{S}_h$ with respect to $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_{i1},\boldsymbol{X}_{i2}}$ as

$$\widetilde{\mathcal{S}}_h = \{\eta(\boldsymbol{x}_2)f_1(x_1)|\delta(x_1,\boldsymbol{x}_2)| \big| f_1\in\widetilde{\mathcal{F}}_h\}.$$

For a function $s_z = \eta(\boldsymbol{X}_2)h^{-1}K\big(h^{-1}(x_1-z)\big)|\delta(x_1,\boldsymbol{x}_2)| \in \mathcal{S}_h$, let

$$s_{\widetilde{z}} = \eta(\boldsymbol{x}_2)h^{-1}K\big(h^{-1}(x_1-\widetilde{z})\big)|\delta(x_1,\boldsymbol{x}_2)| \in \widetilde{\mathcal{S}}_h$$

be the corresponding element in the cover, that is $h^{-1}K\big(h^{-1}(x_1-\widetilde{z})\big) \in \widetilde{\mathcal{F}}_h$ is the corresponding element in the cover for $h^{-1}K\big(h^{-1}(x_1-z)\big) \in \mathcal{F}_h$. Then

$$\begin{aligned}\|s_z - s_{\widetilde{z}}\|^2_{L^2(\mathbb{P}_n)} &= n^{-1}\sum_{i=1}^n \big[\big(K_h(X_{i1}-z) - K_h(X_{i1}-\widetilde{z})\big)\eta(\boldsymbol{X}_{i2})|\delta(X_{i1},\boldsymbol{X}_{i2})|\big]^2 \\ &\le n^{-1}\sum_{i=1}^n \big[K_h(X_{i1}-z) - K_h(X_{i1}-\widetilde{z})\big]^2\|\eta\|^2_\infty r_a^2 \le \epsilon^2.\end{aligned}$$

Therefore, the covering number can be bounded by

$$N\left(\mathcal{S}_h, L^2(\mathbb{P}_n), \epsilon\right) \le N\left(\mathcal{F}_h, L^2(\widehat{\mathbb{P}}_{X_1}), \|\eta\|_\infty^{-1}r_a^{-1}\epsilon\right) \le \left(\frac{2\|K\|_{\mathrm{TV}}A\|\eta\|_\infty r_a}{h\epsilon}\right)^4. \tag{A.3}$$

The variance of the process

$$\sigma_{\mathrm{P}}^2 := \sup_{z\in\mathcal{X}} \mathbb{E}\Big[\big(\eta(\boldsymbol{X}_{12})K_h(X_{11}-z)|\delta_1|\big)^2\Big] \le \|\eta\|^2_\infty r_a^2\mathbb{E}[K_h^2] = \mathcal{O}(r_a^2h^{-1}).$$

Applying Lemma C.1, we have

$$\mathbb{E}\left[\sup_{z\in\mathcal{X}} U_n(z)\right] \le Cn^{-1/2}\sigma_{\mathrm{P}}\sqrt{\log\left(\frac{2\|K\|_{\mathrm{TV}}A\|\eta\|_\infty r_a}{h\sigma_{\mathrm{P}}}\right)} = \mathcal{O}\left(r_a\sqrt{\frac{\log(h^{-1})}{nh}}\right), \tag{A.4}$$

where in the last step we use $\sigma_P = \mathcal{O}(r_a h^{-1/2})$. Next, we bound the supremum of the expectation

$$\begin{aligned}\mathbb{E}\big[\eta(\boldsymbol{X}_{12})K_h(X_{11}-z)\delta_1\big] &= \int \eta(\boldsymbol{X}_{12})K_h(X_{11}-z)|\delta_1|\mathrm{d}\mathbb{P} \\ &\le \|\eta\|_\infty r_a \int K_h(x-z)p_{X_1}(x)\mathrm{d}x = \mathcal{O}(r_a),\end{aligned} \tag{A.5}$$

where in the last step we used the assumption that $\|p_{X_1}\|_\infty$ is bounded from above. Using Lemma C.3, we have

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) \ge \mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big] + t\sqrt{2\Big(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big]\Big)} + \frac{2Ut^2}{3}\right) \le \exp(-nt^2), \tag{A.6}$$

where $\sup_{s_z\in\mathcal{S}_h}\|s_z\|_\infty \le U := h^{-1}\|K\|_\infty\|\eta\|_\infty r_a = \mathcal{O}(r_a h^{-1})$.

Combining (A.4), (A.5) and (A.6), with $t = \sqrt{\log n/n}$, we have with probability at least $1-1/n$,

$$\begin{aligned}\sup_{z\in\mathcal{X}}\frac{1}{n}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\delta}| &\le \sup_{z\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^n \eta(\boldsymbol{X}_{i2})K_h(X_{i1}-z)|\delta_i| \\ &\le \sup_{z\in\mathcal{X}} U_n(z) + \sup_{z\in\mathcal{X}}\mathbb{E}\big[\eta(\boldsymbol{X}_{12})K_h(X_{11}-z)\delta_1\big] = \mathcal{O}(r_a),\end{aligned}$$

where in the last step we used the assumption that $\log n/(nh) = \mathcal{O}(1)$. □

Proof of Lemma 7.2. We rewrite $\sup_{z\in\mathcal{X}} n^{-1}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\xi}_z|$ as

$$\sup_{z\in\mathcal{X}}\frac{1}{n}\Big|\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})^2\big(f(X_{i1})-f(z)\big)\Big|.$$

Denote the empirical process and the function class of interest here as

$$\begin{aligned}U_{n,\boldsymbol{\xi}}(z) &= \frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})^2\big(f(X_{i1})-f(z)\big) - \mathbb{E}\big[K_h(X_{11}-z)\eta(\boldsymbol{X}_{12})^2\big(f(X_{11})-f(z)\big)\big], \\ \mathcal{S}_{h,\boldsymbol{\xi}} &= \big\{s_z(x_1,x_2) = h^{-1}K\big(h^{-1}(x_1-z)\big)\eta^2(\boldsymbol{x}_2)\big(f(x_1)-f(z)\big)\big|z\in\mathcal{X}\big\}.\end{aligned}$$

For simplicity, we will use the notation $U_n(z)$ and $\mathcal{S}_h$ instead, when it leads to no confusion. We note that $\mathcal{S}_h$ is a subset of $\mathcal{F}_{h,1}\times\mathcal{F}_{h,2}$, where $\mathcal{F}_{h,1} := \big\{h^{-1}K\big(h^{-1}(x_1-z)\big)\eta^2(\boldsymbol{x}_2)\big|z\in\mathcal{X}\big\}$ and $\mathcal{F}_{h,2} := \big\{f(x_1)-f(z)\big|z\in\mathcal{X}\big\}$. A similar argument to that used to obtain (A.3) gives us

$$N(\mathcal{F}_{h,1}, L^2(\mathbb{P}_n), \epsilon) \le \left(\frac{2\|K\|_{\mathrm{TV}}A\|\eta\|_\infty^2}{h\epsilon}\right)^4.$$

For $f \in \mathcal{H}(2, L)$ we have $\big|\big(f(x_1) - f(z_1)\big) - \big(f(x_1) - f(z_2)\big)\big| \le \|f'\|_\infty |z_1 - z_2|$, and

$$N(\mathcal{F}_{h,2}, L^2(\mathbb{P}_n), \epsilon) \le N(\mathcal{X}, d(\cdot,\cdot), \|f'\|_\infty^{-1}\epsilon) \le \frac{\|f'\|_\infty \mathrm{Diam}(\mathcal{X})}{\epsilon}.$$

By Lemma C.7 and Lemma C.8, we obtain

$$N(\mathcal{S}_h, L^2(\mathbb{P}_n), \epsilon) \le \|f'\|_\infty \mathrm{Diam}(\mathcal{X}) \left(\frac{2\|K\|_{\mathrm{TV}} A \|\eta\|_\infty^2}{h}\right)^4 \left(\frac{\|\mathcal{F}_{h,1}\|_\infty \vee \|\mathcal{F}_{h,2}\|_\infty}{\epsilon/2}\right)^5 = O(h^{-9}\epsilon^{-5}).$$

Furthermore, we bound the variance by applying the Taylor expansion as follows

$$\begin{aligned}
\sigma_{\mathrm{P}}^2 &:= \sup_{z\in\mathcal{X}} \mathbb{E}\left[\Big(K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big(f(X_1) - f(z)\big)\Big)^2\right] \\
&\le \sup_{z\in\mathcal{X}} \|\eta\|_\infty^2 h^{-2} \int K^2\big(h^{-1}(x_1 - z)\big)\big(f(x_1) - f(z)\big)^2 p_{X_1}(x)\mathrm{d}x \\
&= \sup_{z\in\mathcal{X}} \|\eta\|_\infty^2 h^{-1} \int K^2(u)\big(f(z + uh) - f(z)\big)^2 p_{X_1}(z + uh)\mathrm{d}u \\
&= \sup_{z\in\mathcal{X}} \|\eta\|_\infty^2 h^{-1} \int K^2(u)\big(f'(z)uh + o(uh)\big)^2\big(p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)\big)\mathrm{d}u \\
&= \sup_{z\in\mathcal{X}}\Big[\big(f'(z)\big)^2 p_{X_1}(z)\Big] \cdot \int u^2 K^2(u)\mathrm{d}u \cdot \|\eta\|_\infty^2 h + o(h) = \mathcal{O}(h).
\end{aligned}$$

Next, we study the uniform upper bound of functions in $\mathcal{S}_h$. Depending on whether $h^{-1}(x_1 - z)$ lies in the support, we have two cases

- if $h^{-1}(x_1 - z) \notin \mathcal{X}$, then $K_h(x_1 - z)\eta^2(\boldsymbol{x}_2)\big(f(x_1) - f(z)\big) = 0$;
- if $h^{-1}(x_1 - z) \in \mathcal{X}$, then by mean value theorem,

$$\big|K_h(x_1 - z)\eta^2(\boldsymbol{x}_2)\big(f(x_1) - f(z)\big)\big| \le h^{-1}\|K\|_\infty \|f'\|_\infty h \cdot \mathrm{Diam}(\mathcal{X}) = \|K\|_\infty \|f'\|_\infty \mathrm{Diam}(\mathcal{X}).$$

Therefore $\sup_{s_z\in\mathcal{S}_h} \|s_z\|_\infty \le U := \|K\|_\infty \|f'\|_\infty \mathrm{Diam}(\mathcal{X}) = \mathcal{O}(1)$. Combining the results above, by Lemma C.1 we have

$$\mathbb{E}\left[\sup_{z\in\mathcal{X}} U_n(z)\right] \le C n^{-1/2} h^{1/2} \sqrt{\log \frac{h^{-9/5}}{h^{1/2}}} = \mathcal{O}\big(\sqrt{n^{-1} h \log h^{-1}}\big). \tag{A.7}$$

Similar to σ_{P}^2, we bound the expectation as follows

$$\begin{aligned}
&\mathbb{E}[K_h(X_1 - z)\eta(\boldsymbol{X}_2)^2\big(f(X_1) - f(z)\big)] \\
&\quad = h^{-1} \int K\big(h^{-1}(x - z)\big)\big(f(x) - f(z)\big) p_{X_1}(x) \cdot \mathbb{E}[\eta(\boldsymbol{X}_2)^2 | X_1 = x]\mathrm{d}x \\
&\quad = \int K(u)\big[f'(z)uh + f''(z)(uh)^2/2 + o(u^2h^2)\big]\big[p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)\big] \\
&\qquad \cdot \left(\mathbb{E}[\eta(\boldsymbol{X}_2)^2 | X_1 = z] + uh\frac{\mathrm{d}}{\mathrm{d}z}\mathbb{E}[\eta(\boldsymbol{X}_2)^2 | X_1 = z] + o(uh)\right)\mathrm{d}u = \mathcal{O}(h^2),
\end{aligned} \tag{A.8}$$

where in the last step we used the fact that $K(\cdot)$ is an even function and $f \in \mathcal{H}(2, L)$ on a bounded set $\mathcal{X}$. Using Lemma C.3, we have

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) - \mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big] \geq t\sqrt{2\Big(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big]\Big)} + \frac{2Ut^2}{3}\right) \leq \exp(-nt^2), \tag{A.9}$$

where $\sup_{s_z\in\mathcal{S}_h} \|s_z\|_\infty \leq U := \|K\|_\infty \|f'\|_\infty \mathrm{Diam}(\mathcal{X}) = \mathcal{O}(1)$.

Combining (A.7), (A.8), and (A.9) with $t = \sqrt{\log n/n}$, we have with probability at least $1 - 1/n$,

$$\begin{aligned}
\sup_{z\in\mathcal{X}} n^{-1}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\xi}_z| &= \sup_{z\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})^2\big(f(X_{i1}) - f(z)\big) \\
&= \sup_{z\in\mathcal{X}} U_n(z) + \sup_{z\in\mathcal{X}} \mathbb{E}[K_h(X_1-z)\eta(\boldsymbol{X}_2)^2\big(f(X_1)-f(z)\big)] \\
&= \mathcal{O}\big(h^2 + \sqrt{n^{-1}h\log n}\big),
\end{aligned}$$

where in the last step we used the assumption that $(nh)^{-1}\log n = \mathcal{O}(1)$. □

Proof of Lemma 7.3. We denote the empirical process and function class of interest as

$$\begin{aligned}
U_{n,\boldsymbol{\varepsilon}}(z) &= \frac{1}{n}\sum_{i=1}^n \eta(\boldsymbol{X}_{i2})K_h(X_{i1}-z)\varepsilon_i, \\
\mathcal{S}_{h,\boldsymbol{\varepsilon}} &= \{g_z(x_1,\boldsymbol{x}_2) = h^{-1}K(h^{-1}(x_1-z))\eta(\boldsymbol{x}_2) | z\in\mathcal{X}\}.
\end{aligned}$$

For simplicity we will use the notation $U_n(z)$ and $\mathcal{S}_h$ instead, when it leads to no confusion. Since ε_i's are subgaussian, we have $\mathbb{P}(\max_i|\varepsilon_i| \geq C\sqrt{\log n}) \leq 1/n$. In the following, we condition on the event $\mathcal{A} = \big\{\max_i|\varepsilon_i| < C\sqrt{\log n}\big\}$.

Using an argument similar to that in Lemma 7.1 and 7.2, Lemma C.2 gives us

$$N(\mathcal{S}_h, L^2(\mathbb{P}_n), \epsilon) \leq \left(\frac{2\|K\|_{\mathrm{TV}}A\|h\|_\infty}{h\epsilon}\right)^4.$$

We upper bound the variance of the process as

$$\sigma_{\mathrm{P}}^2 := \sup_{z\in\mathcal{X}} \mathbb{E}\left[K_h^2(X_1-z)\eta^2(\boldsymbol{X}_2)\varepsilon_i^2|\mathcal{A}\right] \leq \log n\|\eta\|_\infty^2 \sup_{z\in\mathcal{X}} \mathbb{E}\left[K_h^2(X_1-z)\right] = \mathcal{O}(h^{-1}\log n).$$

By Lemma C.1, we have

$$\mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)|\mathcal{A}\Big] \leq C\sigma_{\mathrm{P}}n^{-1/2}\sqrt{\log\big(2\|K\|_{\mathrm{TV}}A\|h\|_\infty/(h\sigma_{\mathrm{P}})\big)} = \mathcal{O}\Big(\sqrt{\frac{\log n}{nh}}\Big). \tag{A.10}$$

Using Lemma C.3, we have

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) \geq \mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)|\mathcal{A}\Big] + t\sqrt{2\Big(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)|\mathcal{A}\Big]\Big)} + \frac{2Ut^2}{3}\Big|\mathcal{A}\right) \leq \exp(-nt^2), \tag{A.11}$$

where $\sup_{z\in\mathcal{X}} \|\eta(\boldsymbol{X}_{12})K_h(X_{11}-z)\varepsilon_1|\mathcal{A}\|_\infty \le U := C\sqrt{\log n}h^{-1}\|K\|_\infty\|\eta\|_\infty = \mathcal{O}(\sqrt{\log n}h^{-1})$. Combining (A.10) and (A.11), with $t=\sqrt{\log n/n}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) \ge C\sqrt{\log n/(nh)} + C(nh)^{-1/2}\log n\right)\\
&\le \mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) \ge \mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big|\mathcal{A}\Big] + C(nh)^{-1/2}\log n\Big|\mathcal{A}\right) + P(\mathcal{A}^c) \le 1/n + 1/n = 2/n,
\end{aligned}$$

where we used the assumption that $\log n/(nh) = \mathcal{O}(1)$ and the constant C is large enough. Similarly, we have $\mathbb{P}\big(\sup_z\big(-U_n(z)\big) \ge C(nh)^{-1/2}\log n\big) \le 2/n$. Therefore, with probability at least $1-4/n$,

$$\sup_{z\in\mathcal{X}} n^{-1}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\varepsilon}| = \sup_{z\in\mathcal{X}}\big|U_n(z)\big| \le C(nh)^{-1/2}\log n,$$

which completes the proof. □

Proof of Lemma 7.4. Note that

$$\begin{aligned}
\frac{1}{n}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\zeta}| &= \Big|\frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\boldsymbol{\psi}(X_{i1},\boldsymbol{X}_{i2})^{\mathrm{T}}\big(\boldsymbol{\gamma}^*-\widehat{\boldsymbol{\gamma}}\big)\Big|\\
&\le \Big\|\frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\boldsymbol{\psi}(X_{i1},\boldsymbol{X}_{i2})^{\mathrm{T}}\Big\|_{2,\infty}\cdot\big\|\boldsymbol{\gamma}^*-\widehat{\boldsymbol{\gamma}}\big\|_{2,1}.
\end{aligned}$$

We can rewrite the first term in the last equation as

$$\begin{aligned}
\max_{j\in[d]}\sup_{\mathbf{v}\in\mathbb{B}^m} n^{-1}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j(X_{i1},\boldsymbol{X}_{i2}) &\le n^{-1}\sum_{i=1}^n K_h(X_{i1}-z)B^2\sqrt{m}\\
&\le C\sqrt{m},
\end{aligned}$$

where in the last step we use Lemma A.1 and $\|\eta\|_\infty, \|\psi_{jk}\|_\infty \le B$. Thus, with probability $1-c/n$,

$$\frac{1}{n}|\boldsymbol{h}^{\mathrm{T}}\mathbf{W}_z\boldsymbol{\zeta}| \le C\sqrt{m}r_e,$$

which completes the proof. □

B Auxiliary Lemmas for Bootstrap Confidence Bands

We give proofs of technical lemmas stated in Section 4.3. We start by stating with two technical lemmas.

Lemma B.1. Let $P_1(\mathbf{w}) = \mathbb{E}K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\boldsymbol{\psi}(\boldsymbol{X}_i)^{\mathrm{T}}$. If $(nh)^{-1}\log n = \mathcal{O}(1)$, for any fixed $\mathbf{w}$, with $\|\mathbf{w}\|_1 \le L$, there exist a constant C such that

$$\Big\|\frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\boldsymbol{\psi}(\boldsymbol{X}_i)^{\mathrm{T}} - P_1(\mathbf{w})\Big\|_{2,\infty} \le C\cdot\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}},$$

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

Proof. Note that

$$\left\|\frac{1}{n}\sum_{i=1}^{n} K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\boldsymbol{\psi}(\boldsymbol{X}_i)^{\mathrm{T}}-P_1(\mathbf{w})\right\|_{2,\infty}$$
$$=\max_{j\in[d]}\sup_{\mathbf{v}\in\mathbb{B}^m} n^{-1}\sum_{i=1}^{n} K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\ \mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(X_{i1},\boldsymbol{X}_{i2}\right)-\mathbf{v}^T P_1(\mathbf{w}).$$

Let $N_{\mathbf{v}}$ be a $1/2$ -covering of the unit ball $\mathbb{B}^m=\{\mathbf{v}\in\mathbb{R}^m\,|\,\|\mathbf{v}\|_2\le 1\}$ with cardinality $|N_{\mathbf{v}}|\le 6^m$. For any $\mathbf{v}\in\mathbb{B}^m$, there exists $\pi(\mathbf{v})\in N_{\mathbf{v}}$ such that $\|\mathbf{v}-\pi(\mathbf{v})\|_2\le 1/2$. Therefore,

$$\max_{j\in[d]}\sup_{\mathbf{v}\in\mathbb{B}^m}\sum_{i=1}^{n} K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{*\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(\boldsymbol{X}_i\right)-\mathbf{v}^T P_1(\mathbf{w})$$
$$\le 2\max_{j\in[d]}\sup_{\mathbf{v}\in N_{\mathbf{v}}}\sum_{i=1}^{n} K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{*\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(\boldsymbol{X}_i\right)-\mathbf{v}^T P_1(\mathbf{w}).$$

We consider the following process and function class of interest:

$$U_n(j,\mathbf{v},z)=\frac{1}{n}\sum_{i=1}^{n} K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(\boldsymbol{X}_i\right)-\mathbf{v}^{\mathrm{T}}P(\mathbf{w}),$$
$$S_h=\{S_z(x_1,x_2)=h^{-1}K\left(h^{-1}(x_1-z)\right)\left(\eta(x_2)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})\right)\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j(\boldsymbol{x})\,|\,z\in\mathcal{X},j\in[d],\mathbf{v}\in N_{\mathbf{v}}\}.$$

Note that S_h is a subset of $\mathcal{F}_{h,1}\times\mathcal{F}_{h,2}$, where

$$\mathcal{F}_{\eta,1}:=\{h^{-1}K\left(h^{-1}(x_1-z)\right)\left(\eta(x_2)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})\right)|z\in\mathcal{X}\}$$

and

$$\mathcal{F}_{h,2}:=\left\{\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(x_1,\boldsymbol{x}_2\right)\,|\,j\in[d],\mathbf{v}\in N_{\mathbf{v}}\right\}.$$

Since $(\eta(x_2)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x}))\le(L+1)B$, Lemma C.2 gives us

$$N\left(\mathcal{F}_{h,1},L^2\left(\mathbb{P}_n\right),\epsilon\right)\le\left(\frac{2\|K\|_{\mathrm{TV}}A(L+1)B}{h\epsilon}\right)^4.$$

By Lemma C.8 and Lemma C.7, we then obtain

$$N\left(\mathcal{S}_h,L^2\left(\mathbb{P}_n\right),\epsilon\right)\le 6^m d\cdot\left(\frac{2\|K\|_{\mathrm{TV}}A(L+1)B}{h}\right)^4\left(\frac{\sqrt{m}\vee h^{-1}}{\epsilon/2}\right)^4.$$

We upper bound the variance of the process as

$$\sigma_p^2:=\sup_{j,\mathbf{v},z}\mathbb{E}[K_h^2\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)^2\left(\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(\boldsymbol{X}_i\right)\right)^2]$$
$$\le B^4(L+1)^2 m\sup_{z\in\mathcal{X}}\mathbb{E}[K_h^2(X_1-z)]=\mathcal{O}(mh^{-1}).$$

Then, by Lemma C.1, we have

$$\mathbb{E}\left[\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z)\right] \leq C\sigma_{\mathrm{P}} n^{-1/2}\sqrt{\log\left(6^m d\cdot \|K\|_{\mathrm{TV}} A\left(\sqrt{m}\vee h^{-1}\right)/\left(h\sigma_{\mathrm{P}}\right)\right)} = \mathcal{O}\left(\frac{m+\sqrt{m\log\left(dh^{-1}\right)}}{\sqrt{nh}}\right). \quad \text{(B.1)}$$

By Lemma C.3, we further have

$$\mathbb{P}\left(\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z) \geq \mathbb{E}\left[\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z)\right] + t\sqrt{2\left(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\left[\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z)\right]\right) + \frac{2Ut^2}{3}}\right) \leq \exp\left(-nt^2\right), \quad \text{(B.2)}$$

where

$$\sup_{j,\mathbf{v},z} \left\|K_h\left(x_1-z\right)\left(\eta\left(\boldsymbol{x}_2\right)-\mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{x}\right)\right)\mathbf{v}^{\mathrm{T}}\boldsymbol{\psi}_j\left(x_1,\boldsymbol{x}_2\right)\right\|_\infty \leq U := B^2(L+1)\sqrt{m}h^{-1}\|K\|_\infty = \mathcal{O}(\sqrt{m}h^{-1}).$$

Combining (B.1) and (B.2), with $t=\sqrt{\log n/n}$, then, with probability at least $1-1/n$,

$$\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z) \leq \mathbb{E}\left[\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z)\right] + \sqrt{2\sigma_{\mathrm{P}}^2 + 4U\mathbb{E}\left[\sup_{j,\mathbf{v},z} U_n(j,\mathbf{v},z)\right]}\cdot\sqrt{\frac{\log n}{n}} + \frac{2U}{3}\cdot\frac{\log n}{n}$$
$$= \mathcal{O}\left(\frac{m+\sqrt{m\log\left(dh^{-1}\right)}}{\sqrt{nh}} + \sqrt{\frac{m\log n}{nh}}\right) = \mathcal{O}\left(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\right),$$

where we used inequality $\sqrt{xy}\leq 2^{-1}(x+y)$ and the assumption that $(nh)^{-1}\log n = \mathcal{O}(1)$. □

Applying Lemma B.1 to $\mathbf{w}^*$ and $\widetilde{\mathbf{w}}$, we have the following corollaries. We will proof that $\|\widetilde{\mathbf{w}}\|_1 \leq L$ later. Note that $P_1(\mathbf{w}^*)=0$.

Corollary B.2. With probability at least $1-1/n$, we have

$$\sup_{z\in\mathcal{X}}\left\|\frac{1}{n}\sum_{i=1}^n K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{*\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)^{\mathrm{T}}\right\|_{2,\infty} \leq C\left(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\right).$$

Corollary B.3. With probability at least $1-2/n$, we have

$$\sup_{z\in\mathcal{X}}\left\|\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i,z)\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)^{\mathrm{T}} - \frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i',z)\boldsymbol{\psi}\left(\boldsymbol{X}_i'\right)^{\mathrm{T}}\right\|_{2,\infty} \leq C\left(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\right),$$

where $t(\boldsymbol{X}_i,z) := K_h(X_{i1}-z)(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i))$ and $\{\boldsymbol{X}_i'\}$ is another dataset for estimating $\widetilde{\mathbf{w}}$.

We also need the following technical result.

Lemma B.4. Let $P_2(\mathbf{w}) = \mathbb{E}K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(X_{i2})$. If $(nh)^{-1}\log n = \mathcal{O}(1)$, for any fixed $\mathbf{w}$, with $\|\mathbf{w}\|_1 \leq L$, there exist a constant C such that

$$\sup_{z\in\mathcal{X}}\left|\frac{1}{n}K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(X_{i2}) - P_2(\mathbf{w})\right| \leq C\cdot\sqrt{\frac{\log n}{nh}},$$

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

Proof. The process and the function class of interest are:

$$U_n(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(X_{i2}) - P_2(\mathbf{w}),$$
$$\mathcal{S}_h = \{S_z(x_1, \boldsymbol{x}_2) := K_h(x_1 - z)\big(\eta(\boldsymbol{x}_2) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})\big)\eta(\boldsymbol{x}_2)|z \in \mathcal{X}\}.$$

Since $\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(X_{i2})$ is bounded, using Lemma C.2, we have

$$N\left(\mathcal{S}_h, L^2\left(\mathbb{P}_n\right), \epsilon\right) \leq \left(\frac{2\|K\|_{\mathrm{TV}}A(L+1)B^2}{h\epsilon}\right)^4,$$

where A is parameter from Lemma C.2 and B is the upper bound for $\|\eta\|_\infty$ and $\|\boldsymbol{\psi}_{jk}\|_\infty$. We can upper bound the variance of the process as:

$$\sigma_p^2 := \sup_z \mathbb{E}K_h(X_{i1} - z)^2\big(\eta(\boldsymbol{X}_{i2}) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2\eta(X_{i2})^2 = \mathcal{O}(h^{-1}).$$

Therefore, from Lemma C.1,

$$\mathbb{E}\left[\sup_{z\in\mathcal{X}} U_n(z)\right] \leq Cn^{-1/2}\sigma_p\sqrt{\log\frac{2\|K\|_{TV}(L+1)B^2}{h\sigma_p}} = C\frac{\sqrt{\log h^{-1}}}{\sqrt{nh}}. \tag{B.3}$$

Since $\sup_z |K_h(x_1 - z)\big(\eta(\boldsymbol{x}_2) - \mathbf{w}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{x})\big)\eta(\boldsymbol{x}_2)| \leq U := Ch^{-1}$, by Lemma C.3, we have

$$\mathbb{P}\left(\sup_z U_n(z) \geq \mathbb{E}\left[\sup_z U_n(z)\right] + t\sqrt{2\left(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\left[\sup_z U_n(z)\right]\right) + \frac{2Ut^2}{3}}\right) \leq \exp\left(-nt^2\right). \tag{B.4}$$

Combining (B.3) and (B.4), with $t = \sqrt{\frac{\log n}{n}}$, then, with probability at least $1 - \frac{1}{n}$, we have

$$\begin{aligned}\sup_z U_n(z) &\leq \mathbb{E}\left[\sup_z U_n(z)\right] + \sqrt{2\sigma_{\mathrm{P}}^2 + 4U\mathbb{E}\left[\sup_z U_n(z)\right]}\cdot\sqrt{\frac{\log n}{n}} + \frac{2U}{3}\cdot\frac{\log n}{n} \\ &\leq C\sqrt{\frac{\log n}{nh}},\end{aligned}$$

where we used the inequality $\sqrt{xy} \leq 2^{-1}(x + y)$ and the assumption that $(nh)^{-1}\log n = \mathcal{O}(1)$. □

Similarly, we can have two other corollaries:

Corollary B.5. With probability at least $1-\frac{1}{n}$, we have

$$\min_{z\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^{n}K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\mathbf{w}^{*\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(X_{i2})\geq b-C\sqrt{\frac{\log n}{nh}}.$$

Note that in Assumption 4.1, $\mathbb{E}K_h(X_1-z)\eta(\boldsymbol{X}_2)(\eta(\boldsymbol{X}_2)-\boldsymbol{w}^{*T}\boldsymbol{\psi}(X))\geq b$.

Corollary B.6. With probability at least $1-\frac{2}{n}$, we have

$$\sup_{z\in\mathcal{X}}\left|\frac{1}{n}\sum_{i=1}^{n}t(\boldsymbol{X}_i,z)\eta(\boldsymbol{X}_{i2})-\frac{1}{n}\sum_{i=1}^{n}t(\boldsymbol{X}_i',z)\eta(\boldsymbol{X}_{i2}')\right|\leq C\sqrt{\frac{\log n}{nh}},$$

where $t(\boldsymbol{X}_i,z):=K_h(X_{i1}-z)(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i))$ and $\{\boldsymbol{X}_i'\}$ is another dataset for estimating $\widetilde{\mathbf{w}}$.

Now we can proof the Lemma 7.5.

Proof of Lemma 7.5. Combining Corollary B.2 and Corollary B.5, with probability $1-\frac{2}{n}$,

$$\sup_{z\in\mathcal{X}}\left\|\frac{1}{n}\sum_{i=1}^{n}K_h\left(X_{i1}-z\right)\left(\eta\left(\boldsymbol{X}_{i2}\right)-\mathbf{w}^{*\mathrm{T}}\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)\right)\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)^{\mathrm{T}}\right\|_{2,\infty}\leq C\left(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\right).$$

Setting $\lambda=C\left(\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}\right)$, $\mathbf{w}^*$ is a feasible solution for the optimization problem (7.7). Since the loss function is $\|\mathbf{w}\|_1$, we can get that with probability at least $1-\frac{2}{n}$, the solution $\widetilde{\mathbf{w}}$ to problem (7.7) exists, and $\|\widetilde{\mathbf{w}}\|_1\leq\|\mathbf{w}^*\|_1\leq L$. □

Proof of Lemma 7.6. The loss function

$$L_z(\alpha,\gamma)=\frac{1}{2n}\sum_{i=1}^{n}K_n(X_{i1}-z)\big(Y_i-\alpha\cdot\eta(\mathbf{X}_{i2})-\gamma^T\psi(\mathbf{X}_i)\big)^2$$

is a polynomial of degree 2. From the optimization problem (7.8),

$$\begin{aligned}
&\nabla_\alpha L_z(\widetilde{\alpha},\widehat{\gamma})-\widetilde{\mathbf{w}}\nabla_\gamma L_z(\widetilde{\alpha},\widehat{\gamma})\\
&=\nabla_\alpha L_z(\widetilde{\alpha},\gamma^*)+\nabla^2_{\alpha\gamma}L_z(\widetilde{\alpha},\gamma^*)(\widehat{\gamma}-\gamma^*)-\widetilde{\mathbf{w}}\nabla_\gamma L_z(\widetilde{\alpha},\gamma^*)-\widetilde{\mathbf{w}}\nabla^2_{\gamma\gamma}L_z(\widetilde{\alpha},\gamma^*)(\widehat{\gamma}-\gamma^*)\\
&=\nabla_\alpha L_z(\widetilde{\alpha},\gamma^*)-\widetilde{\mathbf{w}}\nabla_\gamma L_z(\widetilde{\alpha},\gamma^*)+\left(\nabla^2_{\alpha\gamma}L_z(\widetilde{\alpha},\gamma^*)-\widetilde{\mathbf{w}}\nabla^2_{\gamma\gamma}L_z(\widetilde{\alpha},\gamma^*)\right)(\widehat{\gamma}-\gamma^*)=0.
\end{aligned}\tag{B.5}$$

Furthermore,

$$\begin{aligned}
\nabla_\alpha L_z(\widetilde{\alpha},\gamma^*)&=\nabla_\alpha L_z(\alpha^*,\gamma^*)+\nabla^2_{\alpha\alpha}L_z(\alpha^*,\gamma^*)(\widetilde{\alpha}-\alpha^*),\\
\widetilde{\mathbf{w}}\nabla_\gamma L_z(\widetilde{\alpha},\gamma^*)&=\widetilde{\mathbf{w}}\nabla_\gamma L_z(\alpha^*,\gamma^*)+\widetilde{\mathbf{w}}\nabla^2_{\gamma\alpha}L_z(\alpha^*,\gamma^*)(\widetilde{\alpha}-\alpha^*).
\end{aligned}\tag{B.6}$$

Combining (B.5) and (B.6), we have

$$(\nabla^2_{\alpha\alpha}L_z-\widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z)(\widetilde{\alpha}-\alpha^*)=(\nabla^2_{\alpha\gamma}L_z-\widetilde{\mathbf{w}}\nabla_{\gamma\gamma})(\gamma^*-\widehat{\gamma})+\widetilde{\mathbf{w}}\nabla_\gamma L_z(\alpha^*,\gamma^*)-\nabla_\alpha L_z(\alpha^*,\gamma^*).$$

Therefore,

$$\widetilde{\alpha}-\alpha^*=\frac{(\nabla^2_{\alpha\gamma}L_z-\widetilde{\mathbf{w}}^T\nabla^2_{\gamma\gamma}L_z)(\gamma^*-\widehat{\gamma})+\widetilde{\mathbf{w}}\nabla_r L_z(\alpha^*,\gamma^*)-\nabla_\alpha L_z}{\nabla^2_{\alpha\alpha}L_z-\widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z}.$$

Note that since $\nabla^2_{\alpha\alpha}L_z-\widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z$ is the schur component of the nonsingular Hessian matrix Σ^*_z, it must be positive. □

Then we give an Lemma to bound the denominator of $\widetilde{\alpha}-\alpha^*$.

Lemma B.7. Under Assumption 4.1 and conditions in Theorem 4.2, with probability $1-\frac{c}{n}$,

$$\min_z \nabla^2_{\alpha\alpha}L_z-\widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z\geq b/2.$$

Proof. In terms of Lemma 7.5, now we already have

$$\|\widetilde{\mathbf{w}}\|_1\leq L, \tag{B.7}$$

$$\sup_{z\in\mathcal{X}}\left\|\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i',z)\boldsymbol{\psi}\left(\boldsymbol{X}_i'\right)^{\mathrm{T}}\right\|_{2,\infty}\leq C\Big(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\Big). \tag{B.8}$$

We are going to prove

$$\min_z \nabla^2_{\alpha\alpha}L_z-\widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z=\min_z\frac{1}{n}\sum_{i=1}^n K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\geq b/2.$$

Apply Lemma B.1 to (B.8), with probability $1-c/n$,

$$\left\|\mathbb{E}K_h(X_1-z)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\boldsymbol{\psi}(\boldsymbol{X})\right\|_{2,\infty}\leq C\Big(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\Big).$$

Since

$$\begin{aligned}&\widetilde{\mathbf{w}}^T\cdot\mathbb{E}K_h(X_1-z)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\boldsymbol{\psi}(\boldsymbol{X})+\mathbb{E}K_h(X_1-z)\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\\&=\left[-\widetilde{\mathbf{w}}^T1\right]\Sigma_z\begin{bmatrix}-\widetilde{\mathbf{w}}\\1\end{bmatrix}\geq b\sqrt{\|\widetilde{\mathbf{w}}\|_2+1}\geq b.\end{aligned}$$

And

$$\begin{aligned}&\widetilde{\mathbf{w}}^T\cdot\mathbb{E}K_h(X_1-z)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\boldsymbol{\psi}(\boldsymbol{X})\\&\leq\|\widetilde{\mathbf{w}}^T\|_{2,1}\cdot\left\|\mathbb{E}K_h(X_1-z)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\boldsymbol{\psi}(\boldsymbol{X})\right\|_{2,\infty}\\&\leq\|\widetilde{\mathbf{w}}\|_1\cdot\left\|\mathbb{E}K_h(X_1-z)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\boldsymbol{\psi}(\boldsymbol{X})\right\|_{2,\infty}\\&\leq C\Big(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\Big).\end{aligned}$$

Therefore, we have with probability $1-c/n$,

$$\min_z\mathbb{E}K_h(X_1-z)\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\geq b-C\Big(\frac{m+\sqrt{m\log\left(dnh^{-1}\right)}}{\sqrt{nh}}\Big).$$

Apply Lemma B.4, under conditions in Theorem 4.2, we have with probability $1-c/n$,

$$\min_z \frac{1}{n}\sum_{i=1}^{n} K_h(X_{i1}-z)\eta(\boldsymbol{X}_{i2})\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big) \geq b - C\Big(\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}\Big) - C\sqrt{\frac{\log n}{nh}} \geq b/2,$$

for sufficient large n.

□

Proof of Lemma 7.7. Recall that

$$I_1 = \frac{(\nabla^2_{\alpha\gamma}L_z - \widetilde{\mathbf{w}}^T\nabla^2_{\gamma\gamma}L_z)(\gamma^* - \widehat{\gamma})}{\nabla^2_{\alpha\alpha}L_z - \widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z}.$$

Under Assumption 4.1 $P(\|\widehat{\gamma}-\gamma^*\|_{2,1} > Cr_e) \leq 1/n$, so we focus on upper bounding the numerator and lower bounding the denominator:

$$\|\nabla^2_{\alpha\gamma}L_z - \widetilde{\mathbf{w}}^T\nabla^2_{\gamma\gamma}L_z\|_{2,\infty} = \Big\|\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)^{\mathrm{T}}\Big\|_{2,\infty},$$

$$\nabla^2_{\alpha\alpha}L_z - \widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z = \frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)\eta(\boldsymbol{X}_{i2}).$$

From the optimization problem (7.7) and Lemma 7.5, we have

$$\|\widetilde{\mathbf{w}}\|_1 \leq L,$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i',z)\boldsymbol{\psi}\left(\boldsymbol{X}_i'\right)^{\mathrm{T}}\right\|_{2,\infty} \leq C\left(\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}\right).$$

Combined with Corollary B.3 and Lemma B.7, we get

$$\left\|\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)\boldsymbol{\psi}\left(\boldsymbol{X}_i\right)^{\mathrm{T}}\right\|_{2,\infty} \leq C\left(\frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}\right),$$

$$\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)\eta(\boldsymbol{X}_{i2}) \geq b/2.$$

Therefore, with probability $1-\frac{7}{n}$, we have

$$I_1 = \frac{(\nabla^2_{\alpha\gamma}L_z - \widetilde{\mathbf{w}}^T\nabla^2_{\gamma\gamma}L_z)(\gamma^* - \widehat{\gamma})}{\nabla^2_{\alpha\alpha}L_z - \widetilde{\mathbf{w}}\nabla^2_{\alpha\gamma}L_z} \leq C\cdot r_e \cdot \frac{m+\sqrt{m\log(dnh^{-1})}}{\sqrt{nh}}.$$

□

Proof of Lemma 7.8. We have the following representation

$$\begin{aligned} I_{21} &= \frac{\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)[(f^*(X_{i1})-f(z))\eta(\boldsymbol{X}_{i2}) + g(\boldsymbol{X}_i) - \gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)]}{p(z)} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)(f^*(X_{i1})-f^*(z))\eta(\boldsymbol{X}_{i2})}{p(z)} + \frac{\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)[g(\boldsymbol{X}_i) - \gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)]}{p(z)}. \end{aligned}$$

From Assumption 4.1, $|g(\boldsymbol{X}_i)-\gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)| \le Cr_a$. Since $t(X_i,z) = K_h(X_{i1}-z)(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i))$, it follows from Lemma A.1 that

$$\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)[g(\boldsymbol{X}_i)-\gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)] \le CBr_a \le Cr_a.$$

From Corollary B.4, we have that $p(z) > b/4$. Therefore, with probability $1-\frac{c}{n}$,

$$\frac{\frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)[g(\boldsymbol{X}_i)-\gamma^{*T}\boldsymbol{\psi}(\boldsymbol{X}_i)]}{p(z)} \le 4Cr_a/b = Cr_a. \tag{B.9}$$

Next, we deal with the first term. We consider the process and function class of interest:

$$U_n(z) = \frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i,z)(f^*(X_{i1})-f^*(z))\eta(\boldsymbol{X}_{i2}) - \mathbb{E}t(\boldsymbol{X}_i,z)(f^*(X_{i1})-f^*(z))\eta(\boldsymbol{X}_{i2}),$$
$$S_h = \{S_z(x_1,\boldsymbol{x}_2) = K_h(x_1-z)\big(\eta(\boldsymbol{x}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{x})\big)\eta(\boldsymbol{x}_2)\big(f^*(x_1)-f^*(z)\big)\big|z\in\mathcal{X}\}.$$

Note that $\mathcal{S}_h$ is a subset of $\mathcal{F}_{h,1}\times\mathcal{F}_{h,2}$, where $\mathcal{F}_{h,1} := \big\{h^{-1}K\big(h^{-1}(x_1-z)\big)h^2(\boldsymbol{x}_2)\big|z\in\mathcal{X}\big\}$ and $\mathcal{F}_{h,2} := \big\{f(x_1)-f(z)\big|z\in\mathcal{X}\big\}$. Similar to the argument used to obtain (A.3), we prove

$$N(\mathcal{F}_{h,1},L^2(\mathbb{P}_n),\epsilon) \le \left(\frac{2\|K\|_{\mathrm{TV}}AB^2(L+1)}{h\epsilon}\right)^4.$$

On the other hand, $f\in\mathcal{H}(2,L)$ on a bounded set $\mathcal{X}$ and is hence Lipschitz. Therefore, we have $\big|\big(f(x_1)-f(z_1)\big)-\big(f(x_1)-f(z_2)\big)\big| \le \|f'\|_\infty|z_1-z_2|$, and

$$N(\mathcal{F}_{h,2},L^2(\mathbb{P}_n),\epsilon) \le N(\mathcal{X},d(\cdot,\cdot),\|f'\|_\infty^{-1}\epsilon) \le \frac{\|f'\|_\infty\mathrm{Diam}(\mathcal{X})}{\epsilon}.$$

By Lemma C.7 and Lemma C.8, we obtain

$$N(\mathcal{S}_h,L^2(\mathbb{P}_n),\epsilon) \le \|f'\|_\infty\mathrm{Diam}(\mathcal{X})\left(\frac{2\|K\|_{\mathrm{TV}}A(L+1)B^2}{h}\right)^4\left(\frac{\|\mathcal{F}_{h,1}\|_\infty\vee\|\mathcal{F}_{h,2}\|_\infty}{\epsilon/2}\right)^5 = O(h^{-9}\epsilon^{-5}).$$

We bound the variance using the Taylor expansion as follows

$$\begin{aligned}
\sigma_{\mathrm{P}}^2 &:= \sup_{z\in\mathcal{X}}\mathbb{E}\left[\Big(K_h(X_1-z)\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2)-\widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X})\big)\big(f(X_1)-f(z)\big)\Big)^2\right]\\
&\le \sup_{z\in\mathcal{X}}(L+1)B^2h^{-1}\int K^2\big(h^{-1}(x_1-z)\big)\big(f(x_1)-f(z)\big)^2p_{X_1}(x)\mathrm{d}x\\
&= \sup_{z\in\mathcal{X}}(L+1)B^2h^{-1}\int K^2\big(u\big)\big(f(z+uh)-f(z)\big)^2p_{X_1}(z+uh)\mathrm{d}u\\
&= \sup_{z\in\mathcal{X}}(L+1)B^2h^{-1}\int K^2\big(u\big)\big(f'(z)uh+o(uh)\big)^2\big(p_{X_1}(z)+p'_{X_1}(z)uh+o(uh)\big)\mathrm{d}u\\
&= \sup_{z\in\mathcal{X}}\Big[\big(f'(z)\big)^2p_{X_1}(z)\Big]\cdot\int u^2K^2\big(u\big)\mathrm{d}u\cdot(L+1)B^2h+o(h) = \mathcal{O}(h).
\end{aligned}$$

Next, we study the uniform upper bound of functions in $\mathcal{S}_h$. Depending on whether $h^{-1}(x_1-z)$ lies in the support, we have two cases

- if $h^{-1}(x_1 - z) \notin \mathcal{X}$, then $K_h(x_1 - z)h^2(\boldsymbol{x}_2)\big(f(x_1) - f(z)\big) = 0$;
- if $h^{-1}(x_1 - z) \in \mathcal{X}$, then by mean value theorem,

$$\begin{aligned}&\big|K_h(x_1 - z)\eta(\boldsymbol{x}_2)\big(\eta(\boldsymbol{x}_2) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{x})\big)\big(f(x_1) - f(z)\big)\big| \\ &\qquad\qquad \le h^{-1}\|K\|_\infty\|f'\|_\infty B^2(L+1)\cdot\mathrm{Diam}(\mathcal{X}) = \|K\|_\infty\|f'\|_\infty\mathrm{Diam}(\mathcal{X}).\end{aligned}$$

Therefore $\sup_{s_z\in\mathcal{S}_h}\|s_z\|_\infty \le U := \|K\|_\infty\|f'\|_\infty\mathrm{Diam}(\mathcal{X}) = \mathcal{O}(1)$. By Lemma C.1, we then have

$$\mathbb{E}\left[\sup_{z\in\mathcal{X}} U_n(z)\right] \le Cn^{-1/2}h^{1/2}\sqrt{\log\frac{h^{-9/5}}{h^{1/2}}} = \mathcal{O}\big(\sqrt{n^{-1}h\log h^{-1}}\big). \tag{B.10}$$

We bound the expectation as follows

$$\begin{aligned}&\mathbb{E}[K_h(X_1 - z)\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\big(f(X_1) - f(z)\big)] \\ &\quad= h^{-1}\int K\big(h^{-1}(x-z)\big)\big(f(x) - f(z)\big)p_{X_1}(x)\cdot\mathbb{E}[\phi(\boldsymbol{X})|X_1 = x]\mathrm{d}x \\ &\quad= \int K(u)\big[f'(z)uh + f''(z)(uh)^2/2 + o(u^2h^2)\big]\big[p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)\big] \\ &\qquad\qquad\cdot\left(\mathbb{E}[\phi(\boldsymbol{X})|X_1 = z] + uh\frac{\mathrm{d}}{\mathrm{d}z}\mathbb{E}[\phi(\boldsymbol{X})|X_1 = z] + o(uh)\right)\mathrm{d}u = \mathcal{O}(h^2),\end{aligned} \tag{B.11}$$

where in the last step we used the fact that $K(\cdot)$ is an even function, $f \in \mathcal{H}(2, L)$ on a bounded set $\mathcal{X}$, and $\phi(\boldsymbol{X}) = \eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)$. Using Lemma C.3, we then have

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}} U_n(z) - \mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big] \ge t\sqrt{2\Big(\sigma_{\mathrm{P}}^2 + 2U\mathbb{E}\Big[\sup_{z\in\mathcal{X}} U_n(z)\Big]\Big) + \frac{2Ut^2}{3}}\right) \le \exp(-nt^2), \tag{B.12}$$

where $\sup_{s_z\in\mathcal{S}_h}\|s_z\|_\infty \le U := \|K\|_\infty\|f'\|_\infty\mathrm{Diam}(\mathcal{X}) = \mathcal{O}(1)$. Combining (B.10), (B.11), and (B.12), with $t = \sqrt{\log n/n}$, we have with probability at least $1 - 1/n$,

$$\begin{aligned}&\sup_{z\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^n K_h(X_{i1} - z))\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\big(f(X_{i1}) - f(z)\big) \\ &\qquad= \sup_{z\in\mathcal{X}} U_n(z) + \sup_{z\in\mathcal{X}}\mathbb{E}[K_h(X_1 - z))\eta(\boldsymbol{X}_2)\big(\eta(\boldsymbol{X}_2) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\big(f(X_1) - f(z)\big)] \\ &\qquad= \mathcal{O}\big(h^2 + \sqrt{n^{-1}h\log n}\big),\end{aligned}$$

where in the last step we used the assumption that $(nh)^{-1}\log n = \mathcal{O}(1)$. Combined with $p(z) > b/4$, we have that, with probability at least $1 - c/n$,

$$\frac{\frac{1}{n}\sum_{i=1}^n t(\boldsymbol{X}_i, z)(f^*(X_{i1}) - f^*(z))\eta(\boldsymbol{X}_{i2})}{p(z)} \le 4C\frac{h^2 + \sqrt{n^{-1}h\log n}}{b} \le C\big(h^2 + \sqrt{n^{-1}h\log n}\big). \tag{B.13}$$

Combining (B.13) and (B.9), we have

$$I_{21} \le C\big(h^2 + \sqrt{n^{-1}h\log n} + r_a\big),$$

with probability $1 - \frac{c}{n}$, which completes the proof. □

We need the following two technical lemmas to prove Lemma 7.9.

Lemma B.8. With probability $1 - \frac{c}{n}$,

$$\left|\widehat{\sigma}^2 - \sigma^2\right| \le Cr_n,$$

where $\widehat{\sigma}^2 := n^{-1}\sum_{i=1}\widehat{\varepsilon}_i^2 = n^{-1}\sum_{i=1}^n \left(Y_i - \widehat{f}(X_{i1})\eta(\boldsymbol{X}_{i2}) - \widehat{g}(X_{i1}, \boldsymbol{X}_{i2})\right)^2$.

Proof. We have the following decomposition

$$\begin{aligned}\left|\widehat{\sigma}^2 - \sigma^2\right| &= \left|\frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i^2 - \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \sigma^2\right| \\ &\le \underbrace{\frac{1}{n}\sum_{i=1}^n (\widehat{\varepsilon}_i - \varepsilon_i)^2}_{\text{(I)}} + \underbrace{\left|\frac{2}{n}\sum_{i=1}^n (\widehat{\varepsilon}_i - \varepsilon_i)\varepsilon_i\right|}_{\text{(II)}} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i - \sigma^2\right|}_{\text{(III)}}.\end{aligned}$$

For all i, applying Theorem 4.1, the following holds with probability $1 - \frac{7}{n}$:

$$\begin{aligned}\left|\widehat{\varepsilon}_i - \varepsilon_i\right| &= \left|Y_i - \varepsilon_i - \widehat{f}(X_{i1})\eta(\boldsymbol{X}_{i2}) - \widehat{g}(\boldsymbol{X}_i)\right| \\ &\le \left|\left(f_*(X_{i1}) - \widehat{f}(X_{i1})\right)\eta(\boldsymbol{X}_{i2})\right| + \left|g(\boldsymbol{X}_i) - \widehat{g}(\boldsymbol{X}_i)\right| \\ &\le \left|\left(f_*(X_{i1}) - \widehat{f}(X_{i1})\right)\eta(\boldsymbol{X}_{i2})\right| + \left|g(\boldsymbol{X}_i) - \gamma^*\boldsymbol{\psi}(\boldsymbol{X}_i)\right| + \left|\widehat{\gamma}\boldsymbol{\psi}(\boldsymbol{X}_i) - \gamma^*\boldsymbol{\psi}(\boldsymbol{X}_i)\right| \\ &\le Cr_n + Cr_a + C\sqrt{m}r_e \\ &\le Cr_n,\end{aligned}$$

where $r_n := r_a + \sqrt{n^{-1}h\log n} + h^2 + (nh)^{-1/2}\log n + r_e\sqrt{m}$. Applying Bernstein's inequality (Bernstein, 1964), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \sigma^2\right| > C_1\sqrt{\frac{\sigma^2\log n}{n}}\right) \le \frac{2}{n} \quad \text{and} \quad \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n |\varepsilon_i| - \mathbb{E}|\varepsilon|\right| > C_2\sqrt{\frac{\sigma^2\log n}{n}}\right) \le \frac{2}{n}.$$

Suppose n is large enough, so that $\mathbb{E}|\varepsilon| \ge \sqrt{\sigma^2\log n/n}$. Then

$$\begin{aligned}&\mathbb{P}\left(\text{II} > C(C_2+1)\mathbb{E}|\varepsilon|r_n\right) \\ &\qquad \le \mathbb{P}\left(\max_{i\in[n]}|\widehat{\varepsilon}_i - \varepsilon_i| > Cr_n\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |\varepsilon_i| > \mathbb{E}|\epsilon_i| + C_2\sqrt{\frac{\sigma^2\log n}{n}}\right) \le \frac{11}{n}.\end{aligned}$$

Combining all the bounds, the proof is complete. □

Lemma B.9. With probability $1 - \frac{c}{n}$, we have

$$\sup_z |\widehat{q}^2(z) - q^2(z)| \le C\sqrt{\frac{\log n}{nh^3}}.$$

Proof. Recall that

$$q^2(z) = \mathbb{E}\left[K_h\left(X_1 - z\right)^2 \left(\eta\left(\boldsymbol{X}_2\right) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{X})\right)^2\right], \tag{B.14}$$

$$\widehat{q}^2(z) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_{i1} - z\right)^2 \left(\eta\left(\boldsymbol{X}_{i2}\right) - \widetilde{\mathbf{w}}(z)^{\mathrm{T}} \boldsymbol{\psi}(\boldsymbol{X_i})\right)^2. \tag{B.15}$$

We consider the following process and the function class of interest:

$$U_n(z) = \widehat{q}^2(z) - q^2(z),$$
$$S_\eta(z) = \{S(x_1, x_2) = K_h(x_1 - z)^2(\eta(x_2) - \widetilde{\mathbf{w}}(z)^T \boldsymbol{\psi}(\boldsymbol{x}))^2\}.$$

Due to Lemma C.2, we have

$$N(S_\eta, L^2(\mathbb{P}_n), \epsilon) \leq \left(\frac{2\|K\|_{\mathrm{TV}} A}{h^2 \epsilon}\right)^4.$$

Similarly, we can bound that:

$$\sigma_p^2 = \mathbb{E}\Big[K_h(x_1 - z)^4 \big(\eta(x_2) - \widetilde{\mathbf{w}}(z)^T \boldsymbol{\psi}(\boldsymbol{x})\big)^4\Big] \leq C h^{-3} \mathbb{E} K_h(x_1 - z) = O(h^{-3}).$$

Thus, from Lemma C.1,

$$\mathbb{E}[\sup_z U_n(z)] \leq C\sigma_p n^{-1/2} \sqrt{\log \frac{2\|K\|_{\mathrm{TV}} A}{h^2 \sigma_p}} = C\sqrt{\frac{\log(h^{-1})}{nh^3}}. \tag{B.16}$$

Besides, the upper bound if $S_\eta(z)$ is $U = O(h^{-2})$, since $(\eta(x_2) - \widetilde{\mathbf{w}}(z)^T \boldsymbol{\psi}(\boldsymbol{x}))^2$ can be bounded by constant $(L+1)^2 B^2$. By Lemma C.3, we have

$$\mathbb{P}\left(\sup_z U_n(z) \geq \mathbb{E}\left[\sup_z U_n(z)\right] + t\sqrt{2\left(\sigma_p^2 + 2U\mathbb{E}\left[\sup_z U_n(z)\right]\right) + \frac{2Ut^2}{3}}\right) \leq \exp\left(-nt^2\right). \tag{B.17}$$

Combining (B.16) and (B.17), with $t = \sqrt{\log n / n}$, we have, with probability at least $1 - 1/n$,

$$\begin{aligned}\sup_z |U_n(z)| &\leq \mathbb{E}\left[\sup_z U_n(z)\right] + \sqrt{\frac{\log n}{n}} \cdot \sqrt{2\left(\sigma_p^2 + 2U\mathbb{E}\left[\sup_z U_n(z)\right]\right)} + \frac{2U}{3} \cdot \frac{\log n}{n} \\ &\leq C\sqrt{\frac{\log(h^{-1})}{nh^3}} + C\sqrt{\frac{\log n}{nh^3}} \leq C\sqrt{\frac{\log n}{nh^3}},\end{aligned}$$

which completes the proof. □

Proof of Lemma 7.9. We first establish a uniform lower bound on $q^2(z)$. From the optimization problem 7.7, we have

$$\nabla^2_{\alpha\alpha} L_z - \mathbf{w}^T \nabla^2_{\alpha,\gamma} L_z = \frac{1}{n}\sum_{i=1}^{n} K_h(X'_{i1} - z)\big(\eta(\boldsymbol{X}'_{i2}) - \widetilde{\mathbf{w}}^T \boldsymbol{\psi}(\boldsymbol{X}'_i)\big)\eta(\boldsymbol{X}'_{i2}) \geq b - C\sqrt{\frac{\log n}{nh}}.$$

Note that we set $\lambda_2 = b - C\sqrt{\frac{\log n}{nh}}$. Applying Lemma B.4, we have

$$\mathbb{E}K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(\boldsymbol{X}_{i2}) \geq b - C\sqrt{\frac{\log n}{nh}}. \tag{B.18}$$

Denote $\varphi(\boldsymbol{X}_i) := \eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)$. We then have

$$\begin{aligned}
q^2(z) &= \mathbb{E}K_h(X_{i1} - z)^2\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)^2 \\
&= \int_{\boldsymbol{x}\in\mathcal{X}^d} \frac{1}{h^2}K\left(\frac{x_1 - z}{h}\right)^2 \varphi(\boldsymbol{x})^2 p_{X_1,\boldsymbol{X}_2}(x_1, \boldsymbol{x}_2)\, dx_1\, d\boldsymbol{x}_2 \\
&= \frac{1}{h}\iint_{u\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}^{d-1}} K(u)^2\varphi(z + hu, \boldsymbol{x}_2)^2 p_{X_1,\boldsymbol{X}_2}(z + hu, \boldsymbol{x}_2)\, du\, d\boldsymbol{x}_2 \\
&\geq \frac{b}{h}\iint_{u\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}^{d-1}} K(u)^2\varphi(z + hu, \boldsymbol{x}_2)^2\, du\, d\boldsymbol{x}_2 \\
&\geq \frac{b\Big(\iint_{u\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}^{d-1}} K(u)\varphi(z + hu, \boldsymbol{x}_2)\, du\, d\boldsymbol{x}_2\Big)^2}{h\iint_{u\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}^{d-1}} 1\, du\, d\boldsymbol{x}_2} \\
&\geq \frac{b}{h\|\mathcal{X}\|^d}\cdot\left(\iint_{u\in\mathcal{X},\boldsymbol{x}_2\in\mathcal{X}^{d-1}} K(u)\varphi(z + hu, \boldsymbol{x}_2)\eta(x_2)p_{X_1,\boldsymbol{X}_2}(z + hu, \boldsymbol{x}_2)/B^2\, du\, d\boldsymbol{x}_2\right)^2 \\
&= \frac{b}{h\|\mathcal{X}\|^d B^4}\cdot\Big(\mathbb{E}K_h(X_{i1} - z)\big(\eta(\boldsymbol{X}_{i2}) - \widetilde{\mathbf{w}}^T\boldsymbol{\psi}(\boldsymbol{X}_i)\big)\eta(\boldsymbol{X}_{i2})\Big)^2 \\
&\geq \frac{b}{h\|\mathcal{X}\|^d B^4}\cdot\left(b - C\sqrt{\frac{\log n}{nh}}\right)^2 \\
&\geq C\cdot h^{-1},
\end{aligned}$$

where in the third equation we set $x_1 = z + hu$, in the first inequality we used $p_{X_1,\boldsymbol{X}_2}(z + hu, \boldsymbol{x}_2) \geq b$, in the second inequality we applied Cauchy inequality, in the third inequality we used $\eta(x_2)p_{X_1,\boldsymbol{X}_2}(z + hu, \boldsymbol{x}_2) \leq B^2$ and in the fourth inequality we used (B.18). From Lemma B.9, with probability $1 - \frac{c}{n}$,

$$\sup_z\left|\frac{\widehat{q}^2(z)}{q^2(z)} - 1\right| \leq C\left(\frac{\sqrt{\log n}}{\sqrt{nh}}\right).$$

Since $\left|\frac{\widehat{q}^2(z)}{q^2(z)} - 1\right| = \left|\big(\frac{\widehat{q}(z)}{q(z)} - 1\big)\big(\frac{\widehat{q}(z)}{q(z)} + 1\big)\right|$ and $\frac{\widehat{q}(z)}{q(z)} + 1$ converges to a constant, we have

$$\sup_z\left|\frac{\widehat{q}(z)}{q(z)} - 1\right| \leq C\left(\frac{\sqrt{\log n}}{\sqrt{nh}}\right), \tag{B.19}$$

with probability $1 - \frac{c}{n}$. Similarly, with probability $1 - \frac{7}{n}$,

$$\sup_z\left|\frac{\widehat{\sigma}}{\sigma} - 1\right| \leq Cr_n. \tag{B.20}$$

Finally, since

$$\sup_{z\in\mathcal{X}}\left|\frac{\widehat{\sigma}\widehat{q}(z)}{\sigma q(z)} - 1\right| \leq \sup_{z\in\mathcal{X}}\left|\frac{\widehat{\sigma}}{\sigma} - 1\right| + \sup_{z\in\mathcal{X}}\left|\frac{\widehat{q}(z)}{q(z)} - 1\right|\cdot\left|\frac{\widehat{\sigma}}{\sigma}\right|,$$

with probability $1-\frac{c}{n}$,

$$\sup_{z\in\mathcal{X}}\left|\frac{\widehat{\sigma}\widehat{q}(z)}{\sigma q(z)}-1\right|\leq C\left(\frac{\sqrt{\log n}}{\sqrt{nh}}+r_n\right),$$

which completes the proof. □

Finally, we prove Lemma 7.10, Lemma 7.11 and Lemma 7.12. Recall that

$$\begin{aligned}
\mathbb{G}_n(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i K_h(X_{i1}-z)\big(\eta(X_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\sigma q(z)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i t(\boldsymbol{X}_i,z)}{\sigma q(z)},\\
\mathbb{G}_n^{(1)}(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Bigg[\frac{\varepsilon_i I(|\varepsilon_i|\leq b_n)}{\sigma}\cdot\frac{K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{q(z)}\\
&\qquad -\mathbb{E}\Big(\frac{\varepsilon_i I(|\varepsilon_i|\leq b_n)}{\sigma}\cdot\frac{K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{q(z)}\Big)\Bigg],\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\varepsilon_i I(|\varepsilon|\leq b_n|)t(\boldsymbol{X}_i,z)}{\sigma q(z)}-\mathbb{E}\left(\frac{\varepsilon_i I(|\varepsilon|\leq b_n|)t(\boldsymbol{X}_i,z)}{\sigma q(z)}\right)\right],\\
\mathbb{G}_n^{(2)}(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\xi_i\varepsilon_i I(|\varepsilon_i|\leq b_n)\cdot K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}^{T}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\sigma q(z)}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\xi_i\varepsilon_i I(|\varepsilon_i|\leq b_n)t(\boldsymbol{X}_i,z)}{\sigma q(z)},\\
\widehat{\mathbb{G}}_n(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i\cdot\frac{\widehat{\varepsilon}_i K_h(X_{i1}-z)\big(\eta(\boldsymbol{X}_{i2})-\widetilde{\mathbf{w}}(z)^{\mathrm{T}}\boldsymbol{\psi}(\boldsymbol{X}_i)\big)}{\widehat{\sigma}\widehat{q}(z)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\xi_i\widehat{\varepsilon}_i t(\boldsymbol{X}_i,z)}{\widehat{\sigma}\widehat{q}(z)},
\end{aligned}$$

and

$$W_n=\sup_{z\in\mathcal{X}}\mathbb{G}_n(z),\quad W_n^{(1)}=\sup_{z\in\mathcal{X}}\mathbb{G}_n^{(1)}(z),\quad W_n^{(2)}(j_n)=\sup_{z\in\mathcal{X}}\mathbb{G}_n^{(2)}(z)\ \text{ and }\ \widehat{W}_n(j_n)=\sup_{z\in\mathcal{X}}\mathbb{G}_n(z).$$

Proof of Lemma 7.10. Since ε_i are subgaussian, with probability $1-\frac{2}{n}$,

$$\max_i|\varepsilon_i|\leq b_n.$$

Thus, with probability $1-\frac{2}{n}$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i t(\boldsymbol{X}_i,z)}{\sigma q(z)}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\varepsilon_i I(|\varepsilon_i|\leq b_n|)t(\boldsymbol{X}_i,z)}{\sigma q(z)}. \tag{B.21}$$

Therefore, we just have to bound

$$\sqrt{n}\mathbb{E}\left(\frac{\varepsilon_i I(|\varepsilon_i|\leq b_n|)t(\boldsymbol{X}_i,z)}{\sigma q(z)}\right).$$

For $\mathbb{E}\varepsilon_i I(|\varepsilon_i| \le b_n)$, we have

$$\begin{aligned}
\big|\mathbb{E}\varepsilon_i I(|\varepsilon_i| \le b_n)\big| &= \big|\mathbb{E}\varepsilon - \mathbb{E}\varepsilon_i I(|\varepsilon_i| > b_n)\big| \\
&= \big|\mathbb{E}\varepsilon_i I(|\varepsilon_i| > b_n)\big| \\
&\le \mathbb{E}|\varepsilon_i| \cdot I(|\varepsilon_i| > b_n) \\
&= \int_0^\infty \mathbb{P}\Big(|\varepsilon_i| \cdot I(|\varepsilon_i| > b_n) \ge t\Big) dt \\
&= b_n \cdot P\big(|\varepsilon_i| \ge b_n\big) + \int_{b_n}^\infty \mathbb{P}(|\varepsilon_i| \ge t) dt \\
&\le C\frac{\log n}{n} + \int_{b_n}^\infty e^{-\frac{t^2}{2\sigma^2}} dt \le C\frac{\log n}{n}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\sqrt{n}\mathbb{E}\bigg(\frac{\varepsilon_i I(|\varepsilon_i| \le b_n|)t(\boldsymbol{X}_i, z)}{\sigma q(z)}\bigg) &\le \sqrt{n}\mathbb{E}\bigg(\frac{\big|\varepsilon_i I(|\varepsilon_i| \le b_n|)\big| \cdot \big|t(\boldsymbol{X}_i, z)\big|}{\sigma q(z)}\bigg) \\
&\le C\frac{\log nh^{1/2}}{\sigma n^{1/2}} \cdot \mathbb{E}\big(|t(\boldsymbol{X}_i, z)|\big) \\
&\le C\frac{4(L+1)B^2 \log nh^{1/2}}{n^{1/2}\sigma b} \le C\frac{\log n \cdot h^{1/2}}{n^{1/2}}. \qquad \text{(B.22)}
\end{aligned}$$

Hence, combining (B.21) and (B.22), with probability $1 - \frac{2}{n}$,

$$|W_n - W_n^{(1)}| \le C\frac{\log n \cdot h^{1/2}}{n^{1/2}} \le Cn^{-c},$$

which completes the proof. □

Proof of Lemma 7.11. In this proof, our goal is to bound $W_n^{(1)} - W_n^{(2)}$. Since $W_n^{(1)} = \sup_{z\in\mathcal{X}} \mathbb{G}_n^{(1)}(z)$, $W_n^{(2)}(\boldsymbol{j}_n) = \sup_{z\in\mathcal{X}} \mathbb{G}_n^{(2)}(z)$ and $\mathbb{G}_n^{(1)}(z), \mathbb{G}_n^{(2)}(z)$ have the same denominators, we will focus on the numerators of them. We will define $\mathbb{F}_n^{(1)}$ as the numerator of $\mathbb{G}_n^{(1)}(z)$ and $\mathbb{F}_n^{(2)}$ as a term very close to the numerator of $\mathbb{G}_n^{(2)}(z)$. Firstly, we bound the different between $\mathbb{F}_n^{(1)}$ and $\mathbb{F}_n^{(2)}$. Then we bound the difference between $\mathbb{F}_n^{(2)}$ and the numerator of $\mathbb{G}_n^{(2)}(z)$. Finally, we give the bound of $W_n^{(1)} - W_n^{(2)}(\boldsymbol{j}_n)$.

Let

$$\begin{aligned}
\mathbb{F}_n^{(1)}(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big[\varepsilon_i I(|\varepsilon_i| \le b_n)t(\boldsymbol{X}_i, z) - \mathbb{E}\big(\varepsilon_i I(|\varepsilon_i| \le b_n)t(\boldsymbol{X}_i, z)\big)\Big], \\
\mathbb{F}_n^{(2)}(z) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \cdot \Big[\varepsilon_i I(|\varepsilon_i| \le b_n)t(\boldsymbol{X}_i, z) - \mathbb{E}_n\big(\varepsilon_i I(|\varepsilon_i| \le b_n)t(\boldsymbol{X}_i, z)\big)\Big]
\end{aligned}$$

and

$$M_n^{(1)} = \sup_{z\in\mathcal{X}} \mathbb{F}_n^{(1)}(z) \quad \text{and} \quad M_n^{(2)}(\boldsymbol{J}_n) = \sup_{z\in\mathcal{X}} \mathbb{F}_n^{(2)}(z),$$

where $\boldsymbol{J}_n := \{(\boldsymbol{X}_1, \varepsilon_1), ..., (\boldsymbol{X}_n, \varepsilon_n)\}$. Note that $\mathbb{F}_n^{(1)}(z)$ is a function of $z, \boldsymbol{X}_i, \varepsilon_i$, where $\boldsymbol{X}_i, \varepsilon$ are random variables. However, when analyzing $\mathbb{F}_n^{(2)}(z)$, we consider $\boldsymbol{X}_i, \varepsilon_i$ as fixed samples and the randomness comes from ξ_i. Let

$$\mathcal{S}_h = \{S_z(\epsilon, \boldsymbol{X}) = \epsilon I(|\epsilon| \le b_n)t(\boldsymbol{X}, z) | z \in \mathcal{X}\}.$$

We are going to prove that $\mathcal{S}_h$ is a VC-type class (see Definition C.4). First,

$$\|S_z(\boldsymbol{X},\epsilon)\|_\infty \le b_n\frac{1}{h}(L+1)B^2 = C\frac{\log n}{h} =: b.$$

According to Lemma C.2,

$$N(S_h, L_2(\mathbb{Q}), b\tau) \le \left(\frac{2\|K\|_{tv}A}{\tau}\right)^4,$$

for any finite measure $\mathbb{Q}$. Thus, the parameters (in terms of Definion C.4, there are 4 parameters for VC-type class) for S_h are

$$\begin{cases} v = 4 = \mathcal{O}(1) \\ a = 2\|K\|_{TV}A = \mathcal{O}(1) \\ b = \mathcal{O}(\log n \cdot h^{-1}) \\ \sigma^2 = \mathbb{E}S_z^2(\epsilon,\boldsymbol{X}) = \sigma^2\cdot\dfrac{C}{h} = \mathcal{O}(h^{-1}) \end{cases},$$

and

$$K_h := Av(\log n \vee \log(ab/\sigma)) = O(\log n).$$

Applying Lemma C.5, there exists a tight Gaussian random element $B = \{B(g) : g \in S_h\}$ such that

$$\mathbb{E}\big[B(g_1)B(g_2)\big] = \mathbb{E}\big[g_1(\boldsymbol{X}_1,\varepsilon_1)g_2(\boldsymbol{X}_1,\varepsilon_1)\big] - \mathbb{E}\big[g_1(\boldsymbol{X}_1,\varepsilon_1)\big]\mathbb{E}\big[g_2(\boldsymbol{X}_1,\varepsilon_1)\big],$$

for all $g_1, g_2 \in S_h$. Assume $W^0 := \sup_z B(S_z)$. We have for any $\gamma \in (0,1)$,

$$\mathbb{P}\left(|M_n^{(1)} - W^0| > C\cdot\left(\frac{\log^2 n}{\gamma^{1/2}hn^{1/2}} + \frac{\log^{5/4} n}{\gamma^{1/2}n^{1/4}h^{3/4}} + \frac{\log n}{\gamma^{1/3}n^{1/6}h^{2/3}}\right)\right) \le A'\Big(\gamma + \frac{\log n}{n}\Big),$$

where A' is an absolute constant. Under conditions in Theorem 4.2, since $h \asymp n^{-\delta_2}, \delta_2 < 1/4$, there exists c such that $\frac{1}{n^{1/6}h^{2/3}} \le n^{-c}$. Hence, with $\gamma = n^{-3c/2}$, there exist constants c_1, c_2 such that

$$\mathbb{P}\Big(|M_n^{(1)} - W^0| > c_1 n^{-c}\Big) \le c_2 n^{-c}. \tag{B.23}$$

By applying Lemma C.6 to W^0 and $M_n^{(2)}(\boldsymbol{J}_n)$ we can find that $b^2K_n \le n\sigma^2$ is equivalent to $\frac{\log^3 n}{nh} \le C$ and

$$\phi_n = C\cdot\frac{\log^{1/2} n}{h^{1/2}n^{1/2}} + C\cdot\frac{\log^{5/4} n}{n^{1/4}h^{3/4}} \quad\text{and}\quad \gamma_n(\delta) = C\cdot\frac{1}{\delta}\cdot\frac{\log^{5/4} n}{n^{1/4}h^{3/4}} + \frac{1}{n}.$$

Therefore, if $\frac{\log^3 n}{nh} \le C$, there exists a $S_n \in (\mathcal{X}^d)^n \times [-b_n, b_n]^n$, such that $\mathbb{P}(\boldsymbol{J}_n \in S_n) \ge 1 - \frac{3}{n}$, and for a fixed $\boldsymbol{J}_n = \boldsymbol{j}_n \in S_n$, we have for all $\delta > 0$,

$$\mathbb{P}\Big(\big|M_n^{(2)}(\boldsymbol{j}_n) - W^0\big| > C\cdot\frac{\log^{1/2} n}{h^{1/2}n^{1/2}} + C\cdot\frac{\log^{5/4} n}{n^{1/4}h^{3/4}} + \delta\Big) \le A''\cdot\left(C\cdot\frac{1}{\delta}\cdot\frac{\log^{5/4} n}{n^{1/4}h^{3/4}} + \frac{1}{n}\right),$$

where A'' is an absolute constant. Similarly, under conditions in Theorem 4.2, since $h \asymp n^{-\delta_2}, \delta_2 < 1/4$, there also exists c_1, c_2, c, such that

$$\mathbb{P}\Big(\big|M_n^{(2)}(\boldsymbol{j}_n) - W^0\big| > c_1 n^{-c}\Big) \leq c_2 n^{-c}. \tag{B.24}$$

Combining (B.23) and (B.24), we have for any $\boldsymbol{j}_n \in S_n$,

$$\mathbb{P}\Big(\big|M_n^{(2)}(\boldsymbol{j}_n) - M_n^{(1)}\big| > c_1 n^{-c}\Big) \leq c_2 n^{-c}.$$

Next, we are going to bound:

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\right) \cdot \mathbb{E}_n\big(\varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)\big).$$

Applying Lemma A.1, we have with probability $1 - \frac{1}{n}$,

$$\begin{aligned}\left|\mathbb{E}_n\big(\varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)\big)\right| &= \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)\right| \\ &\leq \frac{1}{n}\sum_{i=1} b_n (L+1) B K_h(X_{i1} - z) \leq C\frac{\log^{3/2} n}{n^{1/2} h^{1/2}}.\end{aligned}$$

Since $\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\right) \sim N(0,1)$, with probability $1 - \frac{2}{n}$,

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\right| \leq C \log n.$$

Hence, with probability $1 - \frac{3}{n}$,

$$\left|\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\right) \cdot \mathbb{E}_n\big(\varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)\big)\right| \leq C\frac{\log^{5/2} n}{n^{1/2} h^{1/2}}.$$

Therefore, under conditions in Theorem 4.2, there exists a set S_n, such that $\mathbb{P}(\boldsymbol{J}_n \in S_n) \geq 1 - \frac{c}{n}$, and for a fixed $\boldsymbol{J}_n = j_n \in S_n$, with probability $1 - \frac{c}{n} - c_2 n^{-c} = 1 - c_2' n^{-c}$,

$$\begin{aligned}|W_n^{(1)} - W_n^{(2)}(j_n)| &\leq \frac{4B}{\sigma b}\left(|M_n^{(1)} - M_n^{(2)}| + \sup_z \left|\mathbb{E}_n\big(\varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)\big)\right|\right) \\ &\leq C \cdot \left(c_1 n^{-c} + \frac{\log^{5/2} n}{n^{1/2} h^{1/2}}\right) \leq c_1' n^{-c}.\end{aligned}$$

□

Proof of lemma 7.12. We introduce an intermediate process:

$$\mathbb{G}_n^{(3)}(z) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\xi_i \varepsilon_i I(|\varepsilon_i| \leq b_n) t(\boldsymbol{X}_i, z)}{\widehat{\sigma}\widehat{q}(z)}.$$

Let $W_n^{(3)}(\boldsymbol{j}_n) = \sup_{z\in\mathcal{X}} \mathbb{G}_n^{(3)}(\boldsymbol{j}_n)$, the big picture of this proof is that we firstly bound $W_n^{(2)}(\boldsymbol{j}_n) - W_n^{(3)}(\boldsymbol{j}_n)$ by some properties of Gaussian process and then we bound $W_n^{(3)}(\boldsymbol{j}_n) - \widehat{W}_n(\boldsymbol{j}_n)$ directly. Combining them, we get the bound for $W_n^{(2)}(\boldsymbol{j}_n) - \widehat{W}_n(\boldsymbol{j}_n)$.

Define

$$\begin{aligned}\Delta\mathbb{G}_n(z) &:= \widehat{\mathbb{G}}_n^{(3)}(z) - \mathbb{G}_n^{(2)}(z) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\xi_i\varepsilon_i I(|\varepsilon_i| \le b_n) t(\boldsymbol{X}_i, z)}{\sigma q(z)} \left(\frac{\sigma q(z)}{\widehat{\sigma}\widehat{q}(z)} - 1\right).\end{aligned}$$

Note that $\Delta\mathbb{G}_n(z)$ is a Gaussian process with mean zero and

$$\mathrm{Var}\big(\Delta\mathbb{G}_n(z)\big) \le \frac{1}{n}\sum_{i=1}^{n} \frac{b_n^2 t^2(\boldsymbol{X}_i, z)}{\sigma^2 q^2(z)} \left(\frac{\sigma q(z)}{\widehat{\sigma}\widehat{q}(z)} - 1\right)^2 \le C\cdot\log^2 n \cdot \left(r_n + \sqrt{\frac{\log n}{nh}}\right)^2 =: \sigma_\Delta, \quad \text{(B.25)}$$

where in the last inequality we use Lemma 7.9, Lemma A.1, $q^2(z) \ge Ch^{-1}$, and $t(\boldsymbol{X}_i, z) \le Ch^{-1}$. Under conditions in Theorem 4.2, there exists $c, \delta > 0$ such that

$$r_n + \sqrt{\frac{\log n}{nh}} < n^{-c-\delta}.$$

Therefore, $\sigma_\Delta < n^{-2c}$. From Lemma 7.9, for a large enough n, there exists a bounded set $S_0 \in \mathcal{X}^d \times R^d$, such that $P\big(\boldsymbol{J}_n = (\boldsymbol{X}_1^n, \varepsilon_1^n) \in S_0\big) \ge 1 - 3/n$ and, for $j_n \in S_0$,

$$\left(\frac{\sigma q(z)}{\widehat{\sigma}\widehat{q}(z)} - 1\right) \le 2.$$

Consider the following function class:

$$\mathcal{K} := \left\{ S_z(\boldsymbol{x}, \varepsilon) = \frac{\varepsilon I(|\varepsilon| \le b_n) t(\boldsymbol{x}, z)}{\sigma q(z)} \left(\frac{\sigma q(z)}{\widehat{\sigma}\widehat{q}(z)} - 1\right) \middle| z \in \mathcal{X}, \boldsymbol{J}_n \in S_0 \right\}.$$

We are going to show that $\mathcal{K}$ is a VC-type class. First,

$$\|S_z(\boldsymbol{x}, \varepsilon)\|_\infty \le C\log n h^{-1/2} K\big(h^{-1}(x_1 - z)\big) \le \log n h^{-1/2} =: b.$$

According to Lemma C.2,

$$N(\mathcal{K}, L_2(\mathbb{Q}), b\tau) \le \left(\frac{2\|K\|_{tv} A}{\tau}\right)^4,$$

for any finite measure $\mathbb{Q}$. Thus the parameters for this VC-type class function are

$$\begin{cases} v = 4 = \mathcal{O}(1) \\ a = 2\|K\|_{TV} A = \mathcal{O}(1) \\ b = \mathcal{O}(\log n \cdot h^{-1/2}) \\ \sigma^2 = \mathbb{E}S_z^2(\boldsymbol{x}, \varepsilon) \le C\mathbb{E}\dfrac{(\log n)^2 K_h(X_{i1} - z)^2}{q^2(z)} = \mathcal{O}(1). \end{cases}$$

In addition,

$$\mathbb{E}\left[\left(\Delta\mathbb{G}_n(z_1)-\Delta\mathbb{G}_n(z_2)\right)^2\right]\leq\mathbb{E}_n\left[\left(S_{z_1}(\boldsymbol{X}_i,\varepsilon_i)-S_{z_2}(\boldsymbol{X}_i,\varepsilon_i)\right)^2\right],$$

for all $z_1,z_2\in\mathcal{X}$. The covering number for the index set $\mathcal{X}$ with respect to the intrinsic semi-metric induced from the Gaussian process $\Delta\mathbb{G}_n(z)$ is bounded by the uniform covering number for the function class $\mathcal{K}$. Therefore, an application of Corollary 2.2.8 in Vaart and Wellner (1997) gives,

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|\right]&\leq C\log^2 n\left(r_n+\sqrt{\frac{\log n}{nh}}\right)^2\sqrt{(1+v)\log\left(\frac{ab}{\log^2 n(r_n+\sqrt{\log n/(nh)^{-1/2}})^2}\right)}\\&\leq C\log^2 n\left(r_n+\sqrt{\frac{\log n}{nh}}\right)^2\sqrt{\log\left(\frac{h^{-1/2}}{\log n(r_n+\sqrt{\log n/(nh)^{-1/2}})^2}\right)}\\&\leq n^{-2c}.\end{aligned}$$

With Borell–TIS inequalityBorell (1975), for all $u>0$,

$$\mathbb{P}\left(\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|\geq\mathbb{E}\left[\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|\right]+u\right)\leq\exp\left(\frac{-u^2}{2\sigma_\Delta^2}\right).$$

Therefore, setting $u=\sqrt{\log n}/n^c$, with probability $1-2/n$,

$$\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|\leq\mathbb{E}\left[\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|\right]+u\leq n^{-c+\delta}.$$

Since $W_n^{(3)}(\boldsymbol{j}_n)=\sup_{z\in\mathcal{X}}\mathbb{G}_n^{(3)}(\boldsymbol{j}_n)$, we have

$$\left|W_n^{(2)}(\boldsymbol{j}_n)-W_n^{(3)}(\boldsymbol{j}_n)\right|\leq\sup_{z\in\mathcal{X}}\left|\Delta\mathbb{G}_n(z)\right|.$$

Hence, for any $\boldsymbol{j}_n\in S_0$,

$$\mathbb{P}\left(\left|W_n^{(2)}(\boldsymbol{j}_n)-W_n^{(3)}(\boldsymbol{j}_n)\right|>n^{-c+\delta}\right)\leq\frac{2}{n}.\tag{B.26}$$

At last, since ε_i are subgaussian, with probability $1-3/n$,

$$\mathbb{G}_n^{(3)}(z)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\xi_i\varepsilon_i I(|\varepsilon_i|\leq b_n)t(\boldsymbol{X}_i,z)}{\widehat{\sigma}\widehat{q}(z)}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\xi_i\varepsilon_i t(\boldsymbol{X}_i,z)}{\widehat{\sigma}\widehat{q}(z)}.$$

Thus,

$$\mathbb{G}_n^{(3)}(z)-\widehat{\mathbb{G}}_n(z)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\xi_i(\varepsilon_i-\widehat{\varepsilon}_i)t(\boldsymbol{X}_i,z)}{\widehat{\sigma}\widehat{q}(z)}.$$

Since, with probability $1-c/n$,

$$\max_i\left|\varepsilon_i-\widehat{\varepsilon}_i\right|\leq Cr_n,$$

we have

$$\max_i |\xi_i| \le C \log n,$$

with probability $1 - 2/n$. Therefore,

$$\sup_z |\mathbb{G}_n^{(3)}(z) - \widehat{\mathbb{G}}_n(z)| \le \sup_z \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\xi_i(\varepsilon_i - \widehat{\varepsilon}_i) t(\boldsymbol{X}_i, z)}{\widehat{\sigma}\widehat{q}(z)} \right|$$
$$\le \sup_z \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{C \log n \cdot r_n |t(\boldsymbol{X}_i, z)|}{\widehat{\sigma}\widehat{q}(z)} \le C \log n \cdot r_n.$$

and

$$\mathbb{P}\Big(\big| W_n^{(3)}(\boldsymbol{j}_n) - \widehat{W}_n(\boldsymbol{j}_n) \big| > C r_n \cdot \log n \Big) \le \frac{c}{n}. \tag{B.27}$$

Combining (B.26) and (B.27), for any $\boldsymbol{j}_n \in S_0$, there exists c_1, c_2, c such that

$$\mathbb{P}\Big(\big| W_n^{(2)}(\boldsymbol{j}_n) - \widehat{W}_n(\boldsymbol{j}_n) \big| > c_1 n^{-c} \Big) \le c_2 n^{-c}.$$

□

C Collection of Known Results on Empirical Processes

In this section, we provide some known results on empirical processes that are used in the technical proofs.

Lemma C.1 (Theorem 3.12, Koltchinskii (2011))**.** Assume that the functions in $\mathcal{F}$, defined on $\mathcal{X}$, are uniformly bounded by a constant U and $F(\cdot)$ is the envelope of $\mathcal{F}$, such that, $|f(x)| \le F(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Define $\sigma_P^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2]$. Let $X_1, \ldots, X_n$ be i.i.d. copies of the random variable X. We denote the empirical measure as $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_{i1}}$. If for some $A, V \ge 0$ and for all $\varepsilon > 0$ and $n \ge 1$, the covering entropy satisfies

$$N(\mathcal{F}, L^2(\mathbb{P}_n); \epsilon) \le \left(\frac{A \|F\|_{L^2(\mathbb{P}_n)}}{\varepsilon} \right)^V,$$

then for any i.i.d. subgaussian mean zero random variables $\varepsilon_1, \ldots, \varepsilon_n$ there exits a universal constant C such that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_{i1}) \right] \le C \left[\sqrt{\frac{V}{n}} \sigma_P \sqrt{\log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P}} + \frac{VU}{n} \log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P} \right].$$

Furthermore, we also have

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X_{i1}) - \mathbb{E}f(X)) \right] \le C \left[\sqrt{\frac{V}{n}} \sigma_P \sqrt{\log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P}} + \frac{VU}{n} \log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P} \right].$$

Lemma C.2 (Lemma 3, Giné and Nickl (2009)). Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a bounded variation function. Define the function class $\mathcal{F}_h = \{K((t-\cdot)/h) \,|\, t \in \mathbb{R}\}$. There exists $A < \infty$ such that for all probability measures Q on $\mathbb{R}$, we have

$$\sup_Q N(\mathcal{F}_h, L^2(Q), \epsilon) \leq \left(\frac{2\|K\|_{\mathrm{TV}} A}{\epsilon}\right)^4, \text{ for any } \epsilon \in (0,1).$$

Lemma C.3 (Bousquet (2002)). Let $X_1, \ldots, X_n$ be independent random variables and $\mathcal{F}$ is a function class such that there exist η_n and τ_n^2 satisfying

$$\sup_{f\in\mathcal{F}} \|f\|_\infty \leq \eta_n \quad \text{and} \quad \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(f(X_{i1})) \leq \tau_n^2.$$

Define the random variable Z being the suprema of an empirical process

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n (f(X_{i1}) - \mathbb{E}f(X_{i1})) \right|.$$

Then for any $z > 0$, we have the following concentration inequality on the suprema

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + z\sqrt{2(\tau_n^2 + 2\eta_n \mathbb{E}Z)} + 2z^2\eta_n/3\right) \leq \exp(-nz^2).$$

Definition C.4 (VC-type class, Chernozhukov et al. (2014a)). Let $\mathcal{G}$ be a class of measurable functions on a measurable space $(S, \mathcal{S})$, and let $b > 0, a \geq e$ and $v \geq 1$ be some constants. Then the class $\mathcal{G}$ is called $\mathrm{VC}(b, a, v)$ type class if it is uniformly bounded in absolute value by b, and the covering numbers of $\mathcal{G}$ satisfy

$$\sup_Q N(\mathcal{G}, L_2(Q), b\tau) \leq (a/\tau)^v, \quad 0 < \tau < 1,$$

where the supremum is taken over all finitely discrete probability measures Q on $(S, \mathcal{S})$.

Lemma C.5 (Slepian–Stein type coupling for suprema of empirical processes, Theorem A.1 in Chernozhukov et al. (2014a)). Let $X_1, ..., X_n$ be i.i.d random variables taking values in a measurable space $(S, \mathcal{S})$. Let $\mathcal{G}$ be a pointwise-measurable $\mathrm{VC}(b, a, v)$ type function class for some $b > 0, a \geq e$ and $v \geq 1$. Let $\sigma^2 > 0$ be any constant such that $\sup_{g\in\mathcal{G}} \mathbb{E}[g(X_1)^2] \leq \sigma^2 \leq b^2$. Define the empirical process

$$\mathbb{G}_n(g) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(g\left(X_i\right) - \mathrm{E}\left[g\left(X_1\right)\right]\right), \quad g \in \mathcal{G},$$

and let

$$W_n := \|\mathbb{G}_n\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |\mathbb{G}_n(g)|$$

denote the supremum of the empirical process. Let $B = \{B(g) \big| g \in \mathcal{G}\}$ be a tight Gaussian random element in $l^\infty(\mathcal{F})$ with mean zero and covariance function

$$\mathbb{E}\left[B\left(g_1\right) B\left(g_2\right)\right] = \mathbb{E}\left[g_1\left(X_1\right) g_2\left(X_1\right)\right] - \mathbb{E}\left[g_1\left(X_1\right)\right] \mathbb{E}\left[g_2\left(X_1\right)\right],$$

for all $g_1, g_2 \in \mathcal{G}$. It is well known that such a process exists under the VC type assumption; see Vaart and Wellner (1997), pages 100-101. For some sufficiently large absolute constant A, let

$$K_n := Av(\log n \vee \log(ab/\sigma)).$$

Then for every $\gamma \in (0,1)$, one can construct a random variable W^0 on an enriched probability space such that: (i) $W^0 \stackrel{d}{=} \|B\|_{\mathcal{G}}$ and (ii)

$$\mathbb{P}\left(|W_n - W^0| > \frac{bK_n}{(\gamma n)^{1/2}} + \frac{(b\sigma)^{1/2}K_n^{3/4}}{\gamma^{1/2}n^{1/4}} + \frac{b^{1/3}\sigma^{2/3}K_n^{2/3}}{\gamma^{1/3}n^{1/6}}\right) \leq A'\left(\gamma + \frac{\log n}{n}\right),$$

where A' is an absolute constant.

Lemma C.6 (Slepian-Stein type coupling for suprema of conditional multiplier processes, Theorem A.2 in Chernozhukov et al. (2014a))**.** Let $\xi_1, ..., \xi_n$ be independent $N(0,1)$ random variables independent of $X_1^n := \{X_1, ..., X_n\}$, and let $\xi_1^n := \{\xi_1, ..., \xi_n\}$. Define the Gaussian multiplier process

$$\widetilde{\mathbb{G}}_n(g) := \widetilde{\mathbb{G}}_n(X_1^n, \xi_1^n) := \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\Big(g(X_i) - \mathbb{E}_n\big[g(X_i)\big]\Big), \quad g \in \mathcal{G},$$

and for $x_1^n \in S^n$. Let $\widetilde{W}_n(x_1^n) := \|\widetilde{\mathbb{G}}_n(x_1^n, \xi_1^n)\|_{\mathcal{G}}$ denote the supremum of this process calculated for a fixed $X_n^n = x_1^n$. In addition, let

$$\phi_n := \sqrt{\frac{\sigma^2 K_n}{n}} + \left(\frac{b^2\sigma^2K_n^3}{n}\right)^{1/4} \quad \text{and} \quad \gamma_n(\delta) := \frac{1}{\delta}\left(\frac{b^2\sigma^2K_n^3}{n}\right)^{1/4} + \frac{1}{n}.$$

Suppose that $b^2K_n \leq n\sigma^2$. Then for every $\delta > 0$, there exist a set $S_{n,0} \in \mathcal{S}^n$ such that $\mathbb{P}(X_1^n \in S_{n,0}) \geq 1 - \frac{3}{n}$ and, for every $x_1^n \in S_{n,0}$, one can construct a random variable W^0 on an enriched probability space such that: $W^0 \stackrel{d}{=} \|B\|_{\mathcal{G}}$ and (ii)

$$\mathbb{P}\Big(|\widetilde{W}_n(x_1^n) - W^0| > (\phi_n + \delta)\Big) \leq A''\gamma_n(\delta),$$

where A'' is an absolute constant.

Lemma C.7. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two function classes defined on $\mathcal{X}$, such that $\mathcal{F}_1 \subseteq \mathcal{F}_2$. It holds that

$$N(\mathcal{F}_1, L_2(Q), \epsilon) \leq N(\mathcal{F}_2, L_2(Q), \epsilon/2),$$

for any measure Q on $\mathcal{X}$.

Proof. For an ϵ-covering set N_2 of $\mathcal{F}_2$ such that $|N_2| = N(\mathcal{F}_2, L_2(Q), \epsilon/2)$. We define an ϵ-covering set N_1 for $\mathcal{F}_1$ as follows: For each $f_2 \in N_2$, we find a function f_1 in $\mathcal{F}_1 \cap \mathbb{B}(f_2, \epsilon/2)$, and all these f_1 forms the set N_1. Next we prove that N_1 constructed in this way is an ϵ-covering of $\mathcal{F}_1$.

By definition of N_2, for each function f' in $\mathcal{F}_1$ we have a function $f_2' \in N_2$ such that $\|f' - f_2'\|_{L^2(Q)} \leq \epsilon/2$. By definition of N_1, for f_2' there exists a function $f_1' \in N_1 \subseteq \mathcal{F}_1$ such that $\|f_2' - f_1'\|_{L^2(Q)} \leq \epsilon/2$. By triangle inequality,

$$\|f' - f_1'\|_{L^2(Q)} \leq \|f' - f_2'\|_{L^2(Q)} + \|f_2' - f_1'\|_{L^2(Q)} \leq \epsilon,$$

and hence N_1 is an ϵ-covering of $\mathcal{F}_1$. □

Lemma C.8. Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two function classes defined on $\mathcal{X}$ satisfying

$$N(\mathcal{F}_1, L_2(Q), a_1\epsilon) \le C_1\epsilon^{-v_1} \text{ and } N(\mathcal{F}_2, L_2(Q), a_2\epsilon) \le C_2\epsilon^{-v_2}$$

for some $C_1, C_2, a_1, a_2, v_1, v_2 > 0$ and any $0 < \epsilon < 1$. Define $\|\mathcal{F}_\ell\|_\infty = \sup\{\|f\|_\infty, f \in \mathcal{F}_\ell\}$ for $\ell = 1, 2$ and $U = \|\mathcal{F}_1\|_\infty \vee \|\mathcal{F}_2\|_\infty$. For the function classes $\mathcal{F}_\times = \{f_1 f_2 \,|\, f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and $\mathcal{F}_+ = \{f_1 + f_2 \,|\, f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, we have for any $\epsilon \in (0, 1)$,

$$N(\mathcal{F}_\times, L_2(Q), \epsilon) \le C_1C_2 \left(\frac{2a_1U}{\epsilon}\right)^{v_1} \left(\frac{2a_2U}{\epsilon}\right)^{v_2},$$
$$N(\mathcal{F}_+, L_2(Q), \epsilon) \le C_1C_2 \left(\frac{2a_1}{\epsilon}\right)^{v_1} \left(\frac{2a_2}{\epsilon}\right)^{v_2},$$

for any measure Q on $\mathcal{X}$.

Proof. For any $\epsilon \in (0, 1)$, let $N_1 = \{f_{11}, \ldots, f_{1N_1}\}$ and $N_2 = \{f_{21}, \ldots, f_{2N_2}\}$ be the $\epsilon/(2U)$-net of $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively with

$$N_1 \le C_1 \left(\frac{2a_1U}{\epsilon}\right)^{v_1} \text{ and } N_2 \le C_2 \left(\frac{2a_2U}{\epsilon}\right)^{v_2}.$$

Define the set $N = \{f_{1j} f_{2k} \,|\, f_{1j} \in N_1, f_{2k} \in N_2\}$. We now show that N is an ϵ-net for $\mathcal{F}_\times$. For any $f_1 f_2 \in \mathcal{F}$, there exist two functions $f_{1j} \in N_1$ and $f_{2k} \in N_2$ such that $\|f_1 - f_{1j}\|_{L_2(Q)} \le \epsilon/(2U)$ and $\|f_2 - f_{2k}\|_{L_2(Q)} \le \epsilon/(2U)$. Moreover, we have $f_{1j} f_{2k} \in N$ and

$$\|f_1 f_2 - f_{1j} f_{2k}\|_{L_2(Q)} \le \|\mathcal{F}_2\|_\infty \|f_1 - f_{1j}\|_{L_2(Q)} + \|\mathcal{F}_1\|_\infty \|f_2 - f_{2k}\|_{L_2(Q)} \le \epsilon.$$

Therefore N is the ϵ-net for $\mathcal{F}_\times$. Similarly, we also have

$$\|(f_1 + f_2) - (f_{1j} + f_{2k})\|_{L_2(Q)} \le \|f_1 - f_{1j}\|_{L_2(Q)} + \|f_2 - f_{2k}\|_{L_2(Q)} \le \epsilon/U.$$

So $N' = \{f_{1j} + f_{2k} \,|\, f_{1j} \in N_1, f_{2k} \in N_2\}$ is the ϵ/U-net of $\mathcal{F}_+$. We finally complete the proof by showing that

$$|N'| = |N| = N_1N_2 \le C_1C_2 \left(\frac{2a_1U}{\epsilon}\right)^{v_1} \left(\frac{2a_2U}{\epsilon}\right)^{v_2}.$$

□