Statistical Inference for Optimal Treatment Regime and Related Problems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Yunan Wu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Lan Wang, Adviser

June 2020

ACKNOWLEDGEMENTS

I'd like to thank my advisor, Professor Lan Wang, for her constant support and patient guidance throughout the past five years. I am very grateful to Lan for discussing intensively with me and providing valuable insights into my dissertation research. I feel so fortunate to have had the opportunity to work with Lan during my academic life.

I would also like to thank my committee members, Professor Yuhong Yang, Professor Snigdhansu Chatterjee, Professor Rui Kuang and Professor Thomas Murray for their inspiring discussions, encouraging comments and valuable questions.

I also thank my collaborator, Dr. Haoda Fu, for providing interesting dataset and inspiring discussions for my research. My sincere thanks also go to my colleagues and friends for their support, help, and advice. They help to make my Ph.D. life productive and enjoyable during the past five years.

Last but not least, I would like to express my gratitude to my parents, Yan Tang and Yang Wu, for their unwavering encouragement and support for my pursuit of PhD.

DEDICATION

To my parents, Yan Tang and Yang Wu.

ABSTRACT

Precision medicine is an innovative practice for disease treatment that takes into account individual variability in genes, environment, and lifestyle for each patient. Its main aim is to estimate and make inference about the optimal treatment regime. Though many successful estimation strategies have been developed, studies on statistical inference have not attracted much attention until recently. In this thesis, we attempt to study several statistical inference problems about the optimal treatment regime and some related problems in precision medicine. My thesis is composed of three parts. In the first part, we follow a nonparametric setup to estimate the optimal treatment regime, and propose a resampling approach for inference. The estimator based on a smoothed value function significantly saves the computational cost, provides adorable theoretical properties, and ensures the validity of resampling procedures. In the second part, we adopt a semiparametric model-assisted approach, and investigate inference about the effect of a group of variables on the optimal decision rule in the high-dimensional setting. Its theoretical properties are rigorously justified, and the proposed algorithm ensures its computational efficiency. The last part introduces a new approach for estimating a high-dimensional error-in-variable regression model. It enjoys the same computational convenience of standard Dantzig estimator in the non-contamination case and requires no additional tuning parameter. Theoretically, we derive its estimation error bound. The computational efficiency of all the proposed estimation and inference procedures are demonstrated by numerical studies.

Contents

List of Tables

Chapter 1

Introduction

The area of precision medicine is of considerable current interest. It is an innovative practice for disease treatment that takes into account individual variability in genes, environment, and lifestyle for each patient. Abundant methods and algorithms have been developed to make the optimal decision for each individual, which is also called *optimal treatment regime*, in recent years. In this thesis, we attempt to study several statistical inference problems about the optimal treatment regime, as well as related problems in precision medicine, and develop some methods and algorithms with solid theoretical foundation for these problems.

1.1 Precision Medicine and Optimal Treatment Regime

Nowadays, there are many treatments derived for complex diseases, such as cancers, heart diseases, and even HIV. In traditional practice of medicine, the "best" treatment for an *average patient* among all candidate treatments is often assigned to all patients with the same disease. However, different patients respond to a treatment differently. An "one-for-all" treatment does not always work well in clinical medicine. The goal of *precision medicine* is to tailor the treatment to an individual based on his/her individual characteristics, such as gender, age, and clinical histories. Nowadays, doctors also wish to involve genetic in-

formation in choosing the best individualized treatment.

In this context, a *treatment regime* is defined as a mapping from the support of covariates to the collection of treatments. Determining the optimal treatment regime to assign the treatment one may benefit most accurately is the main target of precision medicine. In this thesis, we adopt the mean criterion for the optimality. It defines the optimal treatment regime as the one maximizes the average outcome among the population if every individual follows the treatment recommended by this treatment regime.

Substantial efforts have recently been devoted to studying how to estimate the optimal treatment regime given the individual-level information, including Q-learning and A-learning based methods (Watkins and Dayan, 1992; Robins et al., 2000; Murphy, 2003; Moodie and Richardson, 2010; Qian and Murphy, 2011), and classification-based methods (Zhang et al., 2012; Zhao et al., 2012, 2015a; Wang et al., 2018; Qi et al., 2018), among others.

Although there exists a rich literature on estimation, the associated inference problem has not been studied until recently, in particular, the statistical inference about the parameter β_0 indexing the theoretically optimal treatment regime. It aims to quantify the importance of different predictors on making an optimal treatment decision.

Several inference methods have been investigated for Q-learning (Laber et al., 2014; Chakraborty et al., 2013, 2014; Song et al., 2015) and A-learning (Jeng et al., 2018). There are two major blocks in existing methods. One is that all aforementioned inference approaches require reliable model specification, which is often challenging in real data analysis. There are still obstacles to develop inference procedures based on existing robust estimators, see discussions in Section 2.2.2. The other is that existing work mostly deal with the classical asymptotic setting of fixed p and large n, where p is the number of covariates and n is the sample size, and have not addressed the challenge of inference with high-dimensional variables. In precision medicine, high-dimensional inference is important in practice, especially when genetic information is involved. By determining whether

a given subset of covariates is relevant for making the optimal treatment recommendation, researchers can not only identify critical characteristics for the optimal treatment regime, but also save the cost by collecting information more effectively.

In this thesis, we aim to develop inference approaches about optimal treatment regime in precision medicine, which are able to overcome the aforementioned limitations. First, the proposed inference approaches should be robust; to be specific, they should not strongly rely on the model specification. It helps to alleviate the sensitivity of inference with respect to the underlying generative model. In addition, our new approaches are supposed to be applied in high-dimensional analysis, which allow simultaneous inference for a subgroup of covariates.

1.2 High-dimensional Error-in-variables Regression

Besides the statistical inference about optimal treatment regime, there are some additional statistical problems in the area of precision medicine. The *error-in-variables* problem is one of them. It focuses on the issue that some characteristics are measured imprecisely or indirectly, which is common in precision medicine analysis. For example, in clinical trials, sometimes certain important genetic information is unavailable for part of patients due to privacy protection. In such a situation, existing searching methods without any correction may lead to an imprecise result for the optimal regime, especially when the number of characteristics is extremely large. Relevant studies are abundant, see, for example, the comprehensive introduction of Carroll et al. (2006).

In this thesis, we are interested in the more challenging setting where the number of covariates can exceed the sample size. Standard high-dimensional regression procedures such as Lasso (Tibshirani, 1996) and Dantzig selector (Candes and Tao, 2007) can be severely biased if measurement error is ignored and may fail to recover the underlying sparsity pattern. Attention was paid to the high-dimensional measurement error problem only recently. Sev-

eral papers have made important contributions, such as Loh and Wainwright (2012); Datta and Zou (2017); Belloni et al. (2017); Rosenbaum and Tsybakov (2010). Despite important developments in methodology and theory, existing methods always require additional computational efforts, in particular, one or more additional tuning parameters, comparing with those standard approaches in the settings where the complexity of error-in-variables is absent. The selection of extra tuning parameters always leads to more computational time. Therefore, in this thesis, we attempt to propose a new estimator that enjoys both computational convenience and desirable statistical properties in high dimensional analysis. It implies meaningful applications in precision medicine analysis.

1.3 Thesis Outline

The outline of this dissertation is as follows. In Chapter 2, we follow a nonparametric setup to estimate the optimal treatment regime, and propose a resampling approach for inference. The proposed estimator is based on a smoothed approximation of the empirical value function, which leads to high computational efficiency and adorable theoretical properties. In Section 2.3, we demonstrate a good convergence rate for the smoothed estimator, and rigorously justify the validity of our inference approach based on resampling.

In Chapter 3, we investigate inference about the effect of a group of variables on the optimal decision rule in the high-dimensional setting. We adopt a semiparametric model-assisted approach for optimal decision estimation and inference. The semiparametric structure permits nonlinear interaction effect between the covariates and treatment via an unknown smooth link function, which incorporates many existing models as special cases. We derive simultaneous confidence intervals for inference on a group of variables while allowing the number of covariates to exceed the sample size, and show its validity in Section 3.3. Moreover, we propose a new algorithm for efficient computation in high dimension in Section 3.4.

Next in Chapter 4, we introduce a new approach for estimating a high-dimensional error-in-variable regression model. The proposed new estimator enjoys the computational advantage of classical Dantzig estimator with only one tuning parameter and can be solved by standard linear programming software. We demonstrate its theoretical properties inSection 4.3, and show the computational efficiency in Section 4.4.

Lastly, we conclude this dissertation in Chapter 5 with a discussion of our contributions and future extensions of our work.

Chapter 2

Resampling-based Confidence Intervals for Model-free Robust Inference on Optimal Treatment Regimes

2.1 Introduction

Applications in clinical medicine, public policy, internet marketing and other scientific areas often involve seeking for an individualized treatment rule (or regime, policy) to maximize the potential benefit. For example, Gail and Simon (1985) and Zhang et al. (2012) observed that younger patients with primary operable breast cancer and lower PR levels are likely to benefit more from the treatment L-phenylalanine mustard and 5-fluorouracil (PF) rather than from PF plus tamoxifen (PFT). Several successful estimation strategies have been developed, including Q-learning (Watkins and Dayan, 1992; Murphy, 2005b; Chakraborty et al., 2010; Qian and Murphy, 2011; Song et al., 2015), A-learning (Robins et al., 2000; Murphy, 2003, 2005a; Moodie and Richardson, 2010; Shi et al., 2018), model-free methods (Robins et al., 2008; Orellana et al., 2010; Zhang et al., 2012; Zhao et al., 2012, 2015a; Athey and Wager, 2017; Linn et al., 2017; Zhou et al., 2017; Zhu et al., 2017; Wang et al., 2018; Qi et al., 2018; Lou et al., 2018), tree or list-based methods (Laber and Zhao, 2015; Cui et al., 2017; Zhu et al., 2017; Zhang et al., 2018), targeted learning

ensembles approach (Díaz et al., 2018), among others.

Although there exists a rich literature on estimation, the associated inference problem has not been studied until recently. In this setting, there are two separate but related inference targets: one is the parameter $\boldsymbol{\beta}_0$ indexing the theoretically optimal treatment regime and the other is the theoretically optimal value function $V(\boldsymbol{\beta}_0)$. The former inference problem aims to quantify the importance of different predictors on making an optimal treatment decision, while the latter constructs a confidence interval for the maximally achievable expected performance which can be used as a gold standard to evaluate alternative treatment regimes.

For Q-learning, several inference methods have been investigated. Laber et al. (2014) proposed a novel locally consistent adaptive confidence interval for $\boldsymbol{\beta}_0$, Chakraborty et al. (2013) proposed a practically convenient adaptive m-out-of-n bootstrap for inference on $\boldsymbol{\beta}_0$, Chakraborty et al. (2014) introduced a double bootstrap approach for inference for $V(\boldsymbol{\beta}_0)$, Song et al. (2015) considered inference for $\boldsymbol{\beta}_0$ based on the asymptotic distribution theory for penalized Q-learning. Recently, Jeng et al. (2018) developed Lasso-based procedure for inference on $\boldsymbol{\beta}_0$ in the A-learning framework. However, accurate inference based on Q-learning and A-learning needs reliable model specification. Luedtke and Van Der Laan (2016) developed interesting theory for inference for $V(\boldsymbol{\beta}_0)$ under exceptional laws. Their approach requires to estimate the conditional treatment effect either based on a working model or in a completely nonparametric fashion.

Different from current state-of-the-art methods which are mostly model-based, we aim to develop a model-free approach for making inference for both $\boldsymbol{\beta}_0$ and $V(\boldsymbol{\beta}_0)$. This would be useful to alleviate the sensitivity of inference with respect to the underlying generative model, the specification of which is often challenging in real data analysis.

It is known that the parameter indexing the optimal treatment regime $\boldsymbol{\beta}_0$ corresponds to the parameter of the Bayes rule of a weighted classification problem (Qian and Murphy, 2011; Zhang et al., 2012; Zhao et al., 2012). A substantial challenge in inference for $\boldsymbol{\beta}_0$

lies in the nonsmoothness of the decision function. A popular approach is to replaces the 0-1 loss by a computationally convenient surrogate loss such as the hinge loss (Zhao et al., 2012; Zhou et al., 2017; Lou et al., 2018) or the logistic loss (Jiang et al., 2019). Existing theory on generalization error bound guarantees the predictive performance based on the surrogate loss. Especially, the resulted decision rule is Fisher consistent, that is the sign of the decision function matches that $\text{sign}(\boldsymbol{x}^T\boldsymbol{\beta}_0)$. However, it was known in the classification literature that as a cost of the surrogate loss, it is not guaranteed that the parameters indexing the resulted decision rule is consistent for $\boldsymbol{\beta}_0$, see Lin (2002). On the other hand, robust estimator (Zhang et al., 2012) that directly estimates the Bayes rule has a cubic root convergence rate and a nonnormal limiting distribution, as recently revealed in Wang et al. (2018). Inference is challenging due to the nonstandard asymptotics as naive bootstrap procedure is not consistent. Goldberg et al. (2014) proposed a SoftMax Q-learning approach to alleviate the nonsmoothness problem in Q-learning but have not explore the associated inference theory.

This chapter first proposes a smoothed model-free estimator for the optimal treatment regime and introduce a proximal algorithm which substantially improves both the computational speed and the accuracy. We prove that the smoothed robust estimator has an asymptotic normal distribution and converges to $\boldsymbol{\beta}_0$ with a rate that can be made arbitrarily close to $n^{-1/2}$. We then rigorously justify the validity of a resampling approach for inference. Our study focuses on randomized trials. Extension to observational study is discussed in Section 2.6.

The remaining of this chapter is organized as follows. Section 2.2 introduces the new method and algorithm. Section 2.3 carefully studies the statistical properties for estimation and inference. Section 2.4 reports the results from Monte Carlo simulations. Section 2.5 analyzes a clinical data set from the Childhood Adenotonsillectomy Trial (CHAT). Section 2.6 concludes with some discussions. The appendix gives the technical assumptions and presents several useful lemmas. The supplemental file contains additional numerical

results and detailed technical derivations.

2.2 Proposed Methods

2.2.1 Problem Setup

Let A be a binary variable (0 or 1) denoting the treatment. For each subject, we observe a vector of covariates $\boldsymbol{x} \in \mathbb{R}^p$ and an outcome $Y \in \mathbb{R}$. Without loss of generality, we assume that larger outcome is preferred. To evaluate the treatment effect, we adopt the potential or counterfactual outcome framework (Neyman, 1990; Rubin, 1978) for causal inference. Let Y_1^* and Y_0^* be the potential outcome had the subject received treatment 1 and 0, respectively. In reality, we observe either Y_1^* or Y_0^*, but never both. It is assumed that the observed outcome is the potential outcome corresponding to the treatment the subject actually receives (consistency assumption in causal inference), that is $Y = Y_1^* A + Y_0^*(1 - A)$. Assume A and $\{Y_0^*, Y_1^*\}$ are independent conditional on $\boldsymbol{x}$, that is, no unmeasured confounding. In addition, we assume that the stable unit treatment value assumption (Rubin, 1986) and the positivity assumption are both satisfied, where the former requires a subject's outcome from receiving a treatment is not influenced by the treatment received by other subjects and the latter requires that $0 < P(A = a|\boldsymbol{x}) < 1, \forall\, \boldsymbol{x}$, almost surely.

An individualized treatment rule or a treatment regime, denoted by $d(\boldsymbol{x})$, is a mapping from the space of covariates to the set of treatment options $\{0, 1\}$. Let $Y^*(d)$ be the potential outcome had a subject with covariates $\boldsymbol{x}$ received the treatment assigned by $d(\boldsymbol{x})$. We have

$$Y^*(d) = Y_1^* d(\boldsymbol{x}) + Y_0^* \{1 - d(\boldsymbol{x})\}. \tag{2.1}$$

Given a collection $\mathbb{D}$ of treatment regimes, the optimal treatment regime $\arg\max_{d \in \mathbb{D}} \mathrm{E}(Y^*(d))$ leads to the maximal average outcome if being implemented in the population.

In practice, it is often desirable to have an interpretable treatment regime. Here, we

focus on the popular class of index rules, given by $\mathbb{D} = \{\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0) : \boldsymbol{\beta} \in \mathbb{B}\}$, where $\mathrm{I}(\cdot)$ is the indicator function and $\mathbb{B}$ is a compact subset of $\mathbb{R}^p$. For a given $\boldsymbol{\beta} \in \mathbb{B}$, we sometimes write the corresponding treatment regime $\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0)$ as $d_{\boldsymbol{\beta}}(\boldsymbol{x})$ or $d_{\boldsymbol{\beta}}$ for simplicity. The value function $V(\boldsymbol{\beta}) = \mathrm{E}\{Y^*(d_{\boldsymbol{\beta}})\}$ measures the effectiveness of the treatment regime $d_{\boldsymbol{\beta}}$. We are interested in estimating the parameter indexing the optimal rule

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} V(\boldsymbol{\beta}). \tag{2.2}$$

For identifiability, we assume that there exists a covariate with a nonzero coefficient whose conditional distribution given the other covariates is absolutely continuous and its coefficient is normalized to have absolute value one. Without loss of generality (one can rearrange the labels of the predictors), we assume x_1 is a predictor that satisfies the condition. We write $\boldsymbol{\beta} = (\beta_1, \widetilde{\boldsymbol{\beta}}^T)^T \in \mathbb{R}^p$. Correspondingly, we write $\boldsymbol{x} = (x_1, \widetilde{\boldsymbol{x}}^T)^T$. More discussions on alternative identifiability condition can be found in Section 2.6.2.

2.2.2 Challenges of inference based on existing robust estimators

Qian and Murphy (2011), Zhang et al. (2012), Zhao et al. (2012), among other, observed that optimal treatment regime estimation can be reformulated as a weighted classification problem. Specifically, the value function $V(\boldsymbol{\beta})$ can be equivalently expressed as

$$V(\boldsymbol{\beta}) = \mathrm{E}\Big[\frac{Y}{\pi(A,\boldsymbol{x})} I\big\{A = d_{\boldsymbol{\beta}}(\boldsymbol{x})\big\}\Big], \tag{2.3}$$

where $\pi(A,\boldsymbol{x}) = P(A = 1|\boldsymbol{x})$ is the propensity score of the treatment and is equal to 0.5 in a randomized trial. Expression (2.3) is the foundation for robust or policy-search estimators for optimal treatment regime, which aim to alleviate the practical difficulty of specifying a reliable generative regression model.

A robust estimator can be obtained by directly maximizing an unbiased sample esti-

mator of the expectation in (2.3), which was the approach in Zhang et al. (2012). In a randomized trial, based on the observed data $\{(\boldsymbol{x}_i, Y_i, A_i), i = 1, \ldots, n\}$, which are independent copies of $(\boldsymbol{x}, Y, A)$, $V(\boldsymbol{\beta})$ can be consistently estimated by its sample analog

$$V_n(\boldsymbol{\beta}) = \frac{2}{n}\sum_{i=1}^{n}\{A_i \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) + (1 - A_i)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} \leqslant 0)\}Y_i. \tag{2.4}$$

Leaving out the terms in $V_n(\boldsymbol{\beta})$ that do not depend on $\boldsymbol{\beta}$, we can estimate $\boldsymbol{\beta}_0$ by

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} M_n(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n}(2A_i - 1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)Y_i. \tag{2.5}$$

However, as revealed in Wang et al. (2018) such a direct estimator for the Bayes rule belongs to a class of nonstandard M estimators. It converges at a cubic-root rate to a nonnormal limiting distribution that is characterized by the maximizer of a centered Gaussian process with a parabolic drift. The nonstandard asymptotics is a consequence of the so-called *sharp-edge effect* (Kim and Pollard, 1990). Inference based on this approach is challenging due to the nonstandard asymptotics as the naive bootstrap procedure is not consistent. The smoothed estimator we propose alleviates the sharp-edge effect caused by the indicator function and leads to faster convergence rate.

2.2.3 Smoothed Model-free Inference for Optimal Treatment Regime

To facilitate inference, we study an alternative estimator which can be considered as a compromise between the two robust estimation approaches described in Section 2.2.2. For clarity of presentation, we assume that the data are collected from a randomized trial. Instead of replacing the indicator function with the hinge loss function, we replace it with a smoothed approximation. Formally, we estimate $\boldsymbol{\beta}_0$ by

$$\widehat{\boldsymbol{\beta}}_n = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \widetilde{M}_n(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n}(2A_i - 1)K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\Big)Y_i, \tag{2.6}$$

where $K(\cdot)$ is a smoothed approximation to the indicator function, and h_n is a sequence of smoothing parameter that goes to zero as $n \to \infty$. The function $K(\cdot)$ is required to satisfy some general regularity conditions given in Appendix A.2, see also Remark 2.1 in Section 2.3.1.

The motivation for the above new estimator is three-fold. First, as h_n goes to zero at an appropriate rate, the parameter indexing the optimal treatment regime or the Bayes rule can be estimated at a rate arbitrarily close to $n^{-1/2}$, see Section 2.3.1. Second, smoothing the indicator function circumvents the aforementioned nonstandard asymptotics and would lead to a feasible bootstrap inference procedure with theoretical guarantee, see Section 2.3.2. Third, it also alleviates the computational challenge due to nonsmoothness, see Section 2.2.4 for a new efficient algorithm.

For inference, we apply a resampling technique called "weighted bootstrap" which assigns independent and identically distributed positive random weights to each observation. This resampling scheme was proposed in Rubin (1981). Barbe and Bertail (1995) provided a comprehensive introduction, see also Ma and Kosorok (2005) and Cheng and Huang (2010) for recent interesting developments. The bootstrapped estimate of the smoothed robust estimator is defined as

$$\widehat{\boldsymbol{\beta}}_n^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \widetilde{M}_n^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^n r_i(2A_i-1)K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\Big)Y_i, \qquad (2.7)$$

where $r_1, ..., r_n$ are random weights satisfying conditions given in Section 2.3.2. To evaluate the distribution of $\widehat{\boldsymbol{\beta}}_n^*$ in practice, we repeatedly generate independent samples of random weights. Following notation introduced earlier, let $\widehat{\boldsymbol{\beta}}_n^* = (\widehat{\beta}_{n1}^*, \widetilde{\boldsymbol{\beta}}_n^{*T})^T$, where $|\widehat{\beta}_{n1}^*| = 1$ and $\widetilde{\boldsymbol{\beta}}_n^* = (\widehat{\beta}_{n2}^*, \ldots, \widehat{\beta}_{np}^*)^T$. For $j = 2, \ldots, p$, let $\xi_j^{*(\alpha/2)}$ and $\xi_j^{*(1-\alpha/2)}$ be the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantile of the bootstrap distribution of $(nh_n)^{1/2}(\widetilde{\boldsymbol{\beta}}_j^* - \widetilde{\boldsymbol{\beta}}_j)$, respectively, where α is a small positive number. We can estimate $\xi_j^{*(\alpha/2)}$ and $\xi_j^{*(1-\alpha/2)}$ from a large number of bootstrap samples. An asymptotic $100(1-\alpha)\%$ bootstrap confidence interval for β_{0j},

$j = 2, \ldots, p$, is given by

$$\{\widetilde{\boldsymbol{\beta}}_j - (nh_n)^{-1/2}\xi_j^{*(1-\alpha/2)}, \widetilde{\boldsymbol{\beta}}_j - (nh_n)^{-1/2}\xi_j^{*(\alpha/2)}\}. \tag{2.8}$$

Next, we consider inference for the optimal value. Define

$$V_n^*(\boldsymbol{\beta}) = \frac{2}{n}\sum_{i=1}^{n} r_i\{A_i \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) + (1 - A_i)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} \leqslant 0)\}Y_i. \tag{2.9}$$

Note that $V_n^*(\boldsymbol{\beta})$ can be considered as a perturbed version of the V_n defined in (2.4). Let $d^{*(\alpha/2)}$ and $d^{*(1-\alpha/2)}$ be the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantile of the bootstrap distribution of $n^{1/2}\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\}$, respectively. An asymptotic $100(1-\alpha)\%$ bootstrap confidence interval for $V(\boldsymbol{\beta}_0)$ is

$$\{V_n(\widehat{\boldsymbol{\beta}}_n) - n^{-1/2}d^{*(1-\alpha/2)}, V_n(\widehat{\boldsymbol{\beta}}_n) - n^{-1/2}d^{*(\alpha/2)}\}. \tag{2.10}$$

2.2.4 A Proximal Algorithm

The smoothed robust estimator largely alleviates the computational challenge due to the nonsmooth indicator function. However, the objective function is still a nonconvex function of the parameter. Such nonconvexity is inherent to robust estimation of optimal treatment regime (Qian and Murphy, 2011). We employ a proximal gradient descent algorithm, originally proposed in Nesterov (2007), which applies to a large class of nonconvex problems. In our setting, this algorithm substantially improves the computational speed and can accommodate high-dimensional covariates.

Consider an optimization problem with an objective function $\Phi(\boldsymbol{\beta})$. Nesterov (2007) assumes that $\Phi(\boldsymbol{\beta})$ has the decomposition $\Phi(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + \Psi(\boldsymbol{\beta})$, over a convex set Q, where f is a differentiable function but not necessarily convex, and Ψ is closed and convex on Q. In our setting, we take $-\widetilde{M}_n(\boldsymbol{\beta})$ as the f function, and set $\Psi(\boldsymbol{\beta}) \equiv 0$. Following

Nesterov (2007), we generate a sequence of iterates $\{\boldsymbol{\beta}^{(t)}, t = 0, 1, 2, ...\}$ such that

$$\begin{aligned}\boldsymbol{\beta}^{(t)} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{B}}\Big\{ -\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}) - \big\langle\nabla\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}), \boldsymbol{\beta}-\boldsymbol{\beta}^{(t-1)}\big\rangle + \alpha_t\big|\big|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t-1)}\big|\big|^2 + \Psi(\boldsymbol{\beta})\Big\} \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{B}}\Big\{ -\frac{2}{n}\sum_{i=1}^{n}(2A_i-1)K'\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)}}{h_n}\Big)\frac{\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^{t-1})}{h_n}Y_i + \alpha_t\big|\big|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t-1)}\big|\big|^2\Big\},\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product between two vectors. Observe that the above minimization problem has a closed-form solution

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} + (n\alpha_t)^{-1}\sum_{i=1}^{n}(2A_i-1)K'\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^{(t-1)}}{h_n}\Big)\frac{\boldsymbol{x}_i}{h_n}Y_i.$$

Hence the algorithm can be updated efficiently. The algorithm stops when the following criterion is met:

$$\widetilde{M}_n(\boldsymbol{\beta}^{(t)}) < \widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}) + \big\langle\nabla\widetilde{M}_n(\boldsymbol{\beta}^{(t-1)}), \boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\big\rangle - \alpha_t\big|\big|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\big|\big|^2,$$

where α_t is a sequence of small positive numbers. To choose α_t, inspired by Fan et al. (2018), we employ an expanding series, which ensures that the stepsize diminishes during the update process. Details for this algorithm is provided in the supplementary material.

It is worth emphasizing that this algorithm can be easily adapted to the high-dimensional setting by taking $\Psi(\boldsymbol{\beta})$ as a regularization function, such as the L_1 penalty function.

2.3 Statistical Properties

2.3.1 Consistency and Asymptotic Normality of the Smoothed Estimator

To lay the foundation for inference, we first present the statistical properties of the smoothed robust estimator $\widehat{\boldsymbol{\beta}}_n$ defined in (2.6). All the regularity conditions are summarized in Appendix A.2. Theorem 2.1 below shows that $\widehat{\boldsymbol{\beta}}_n$ is consistent for the parameter indexing the optimal treatment regime. Comparing with the asymptotic normality result in Theorem 2.2, the consistency requires very mild conditions and serves as a precursor step for proving asymptotic normality. See Appendix A.5 of the online supplementary material for the proofs of Theorem 2.1 and Theorem 2.2.

Theorem 2.1

Under (A1) - (A3) and assume $K(\cdot)$ satisfies (K1), then $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + o_p(1)$. □

Recall that for identification, we write $\boldsymbol{\beta}_0 = (\beta_{01}, \widetilde{\boldsymbol{\beta}}_0^T)^T \in \mathbb{R}^p$ where $|\beta_{01}| = 1$. Similarly, we write $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_{n1}, \widetilde{\boldsymbol{\beta}}_n^T)^T \in \mathbb{R}^p$ where $|\widehat{\beta}_{n1}| = 1$. With the above consistency result, we have $P(\widehat{\beta}_{n1} = \beta_{01}) \to 1$ as $n \to \infty$. In the following, we focus on studying the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_n$. To this end, we introduce some additional notations. Define $S(z, \widetilde{\boldsymbol{x}}) = \mathrm{E}(Y_1^* - Y_0^* | z, \widetilde{\boldsymbol{x}})$, where $z = \boldsymbol{x}^T\boldsymbol{\beta}_0$. Note that there is a one-to-one transformation between $(z, \widetilde{\boldsymbol{x}})$ and $\boldsymbol{x} = (x_1, \widetilde{\boldsymbol{x}}^T)^T$. Hence, $S(z, \widetilde{\boldsymbol{x}})$ is a measure of the conditional treatment effect. Let $S^{(1)}(z, \widetilde{\boldsymbol{x}})$ denote the partial derivative of $S(z, \widetilde{\boldsymbol{x}})$ with respect to z. Furthermore, we define

$$\boldsymbol{D} = a_1 \mathrm{E}\big\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) \mathrm{E}(Y_1^{*2} + Y_0^{*2} | z = 0,\ \widetilde{\boldsymbol{x}})\big\}, \tag{2.11}$$

$$\boldsymbol{Q} = a_2 \mathrm{E}\big\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) S^{(1)}(0, \widetilde{\boldsymbol{x}})\big\}, \tag{2.12}$$

where $f(z|\widetilde{\boldsymbol{x}})$ denotes the conditional probability density function of z given $\widetilde{\boldsymbol{x}}$, $a_1 =$

$2\int\{K'(\nu)\}^2 d\nu$, and $a_2 = \int \nu K''(\nu) d\nu$, with $K'(\cdot)$ and $K''(\cdot)$ denoting the first- and second-derivative of $K(\cdot)$, respectively. Note that $\boldsymbol{D}$ and $\boldsymbol{Q}$ both depend on unknown functions, e.g., $f(z|\widetilde{\boldsymbol{x}})$, and are complex to approximate analytically. This motivates us to consider a bootstrap approach for inference procedure.

Theorem 2.2

Assume $K(\cdot)$ satisfies (K1) – (K3) for some $b \geqslant 2$, $h_n = o(n^{-1/(2b+1)})$ and $n^{-1}h_n^{-4} = o(1)$. Then under (A1) – (A5),

(1) $\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0) \to N(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-1})$ in distribution as $n \to \infty$.

(2) $\sqrt{n}\{V_n(\widehat{\boldsymbol{\beta}}_n) - V(\boldsymbol{\beta}_0)\} \to N(0, U)$ in distribution as $n \to \infty$, where $V_n(\cdot)$ is defined in (2.4) and $U = \mathrm{Var}\{Y^*(d_{\boldsymbol{\beta}_0})\} + \mathrm{E}\{(Y^*(d_{\boldsymbol{\beta}_0})^2\}$. □

Remark 2.1

Theorem 2.2 implies that $\widetilde{\boldsymbol{\beta}}_n$ achieves a convergence rate arbitrarily close to $n^{-b/(2b+1)}$. The cumulative distribution function of $N(0,1)$ satisfies these regularity conditions with $b = 2$, and would produce a convergence rate arbitrarily close to $n^{-2/5}$. With a carefully designed $K(\cdot)$ function which satisfied (K1) – (K3) with b sufficiently large, the convergence rate can be further improved. For example, $K(v) = \left[0.5 + \frac{105}{64}\{\frac{v}{5} - \frac{5}{3}(\frac{v}{5})^3 + \frac{7}{5}(\frac{v}{5})^5 - \frac{3}{7}(\frac{v}{5})^7\}\right] I(-5 \leqslant v \leqslant 5) + I(v > 5)$ satisfies (K1) – (K3) with $b = 4$. This choice leads to a convergence rate of $n^{-4/9}$. This function first appeared in Horowitz (1992), which dealt with smoothing estimator in a different setting. Our setting and proofs are very different. Especially, our proofs substantially simplified the traditional methods for handling a smoothed objective function. Example 2 in Appendix A.10 of the supplementary material demonstrates that the performance of the smoothed estimator is not sensitive to the choice of $K(\cdot)$ in finite samples. We would recommend the distribution function of N(0,1) as the default choice due to its simplicity, which we observe to have satisfactory performance in a variety of settings. □

Remark 2.2

The key components of the proofs are modern empirical process techniques. In particular, we introduce some recent empirical process results (Giné and Sang, 2010; Mason, 2012) on VC classes of functions that involve smoothing parameters, which were originally developed for uniform asymptotics with data-driven bandwidth selection and have not been applied to the types of problems considered here. These new techniques lead to simpler proof and are of independent interest. Our technical derivation for this and other results in this chapter employ recent techniques developed by Giné and Sang (2010) and Mason (2012) for VC classes of functions that involve smoothing parameters, see Appendix A.2. Carefully handling function classes involving a smoothing parameter is nontrivial. The literature usually either impose a lower positive bound on h to avoid the process to blow up or requires more involved computation on the entropy bound for such classes. In contrast, the new techniques are based on a geometric argument and avoid the usually intensive entropy computation. The asymptotic normality result in part (2) of the theorem is mostly due to the fact the estimated value function $V_n(\boldsymbol{\beta})$ is a sample average of functions that enjoy the Donsker property. Furthermore, the population value function $V(\boldsymbol{\beta})$ has gradient zero at the true value $\boldsymbol{\beta}_0$. □

2.3.2 Justification for Resampling-based Inference

Let $r_1, ..., r_n$ be a random sample from a distribution of a positive random variable with mean one and variance one. Assume the random weights $r_1, ..., r_n$ are independent of the data. Recall that

$$\widehat{\boldsymbol{\beta}}_n^* \;=\; \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \widetilde{M}_n^*(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n} r_i(2A_i-1)K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\Big)Y_i.$$

Hence, two different sources of randomness contribute to the distribution of $\widehat{\boldsymbol{\beta}}_n^*$ in this setup: one due to the random data and the other due to the random weights.

We next provide a rigorous justification for the validity of the bootstrap procedures proposed in Section 2.2.3. We establish that the bootstrap distribution asymptotically imitates the distribution of the original estimator. Let $r = \{r_1, \ldots, r_n\}$ be the collection of the random bootstrap weights and $w = \{W_1, \ldots, W_n\}$ be the random sample of observations, where $W_i = (\boldsymbol{x}_i, A_i, Y_i)$.

Given a sequence of random variables R_n, $n = 1, \ldots, n$, we write $R_n = o_{p_r}(1)$ if for any $\epsilon > 0, \delta > 0$, we have $P_w(P_{r|w}(|R_n| > \epsilon) > \delta) \to 0$ as $n \to \infty$. In the bootstrap literature, R_n is said to converge to zero in probability, conditional on the data.

Theorem 2.3

Under (A1) – (A3), (A6) and assume $K(\cdot)$ satisfies (K1), then

(1) $\widehat{\boldsymbol{\beta}}_n^* = \widehat{\boldsymbol{\beta}}_n + o_{p_r}(1)$;

(2) $\sqrt{n}\big\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\big\} = N(0, U) + o_{p_r}(1)$. □

Part (2) of Theorem 2.3 suggests that we can use the perturbed value function defined in (2.9) with the plugged-in estimator $\widehat{\boldsymbol{\beta}}_n$ to estimate the asymptotic variance of the estimated optimal value in Theorem 2.2. This establishes the asymptotic validity of the confidence interval in (2.10), which allows for inference for the value function. The validity of the confidence interval in (2.8) for $\boldsymbol{\beta}_0$ is ensured by Theorem 2.4 below.

Theorem 2.4

Assume $K(\cdot)$ satisfies (K1) – (K3) for some $b \geqslant 2$, $h_n = o(n^{-1/(2b+1)})$, and $\log(n) = o(nh_n^4)$. Under (A1) - (A6), $\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n^* - \widetilde{\boldsymbol{\beta}}_n) = N(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{D}\boldsymbol{Q}^{-1}) + o_{p_r}(1)$. □

Remark 2.3

The proofs of Theorems 2.3 and 2.4 are given in Appendix A.5 of the online supplementary material. We make use of the recent results in which allow for using an unconditional argument to derive conditional results. The use of the unconditional argument can be particularly convenient to combine with the Donsker class properties. □

To better understand the behavior of the proposed inference procedure, we also study the properties of the smoothed estimator and its bootstrapped version under a moving parameter or local asymptotic framework. See Appendix A.7 of the online supplementary material.

2.4 Simulation Results

We generate random data from the model $Y = \exp(\boldsymbol{x}^T\boldsymbol{\eta}) + A\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim N(0,1)$, $\boldsymbol{x} = (x_0, x_1, x_2, x_3)^T = (x_0, \widetilde{\boldsymbol{x}}^T)^T$, $x_0 = 1$ and $\widetilde{\boldsymbol{x}}$ follows a 3-dimensional multivariate normal distribution with mean zero and identity covariance matrix. We set $\boldsymbol{\eta} = (-1, -0.5, 0.5, -0.5)^T$, and consider two settings for $\boldsymbol{\beta}$. In setting 1, we have $\boldsymbol{\beta} = (-2, -2, 2, 2)^T$; while in setting 2 we have $\boldsymbol{\beta} = (-2, -2, 2, 0)^T$ with x_3 being an inactive variable for the optimal treatment regime. The optimal treatment regime is given by $\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} \leqslant 0)$. As discussed in Section 2.2.1, for identifiability, we adopt the normalization $|\beta_1| = 1$, corresponding to the coefficient of the continuous covariate x_1. Under this normalization, the population parameter indexing the optimal treatment regime is $\boldsymbol{\beta}^{opt} = (\beta_0^{opt}, \beta_1^{opt}, \beta_2^{opt}, \beta_3^{opt}) = (-1, -1, 1, 1)$ in setting 1, and $(-1, -1, 1, 0)$ in setting 2. We consider 1000 simulation runs and three different sample sizes $n = 300,\ 500,\ 1000$ in the simulation experiments. The confidence intervals are constructed based on 500 bootstrap estimates for each simulation run. That is, for each simulation run, we generate 500 independent samples of size n of positive random weights from a distribution with mean one and variance one and apply them to weight the original observations according to (2.7).

We first study the finite sample performance of the smoothed robust estimator in Section 2.2.3. The smoothed robust estimator is computed using the proximal algorithm in Section 2.2.4, where we choose $K(\cdot)$ to be the cumulative distribution function of standard normal distribution and set $h_n = 0.9n^{-0.2}\min\{\ \mathrm{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}),\ \mathrm{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta})/1.34\}$, as suggested in Silverman (1986), where "std" denotes the standard deviation function, and "IQR" de-

notes the interquartile range. The initial estimator $\boldsymbol{\beta}^0$ in the proximal algorithm is set as $(0, \cdots, 0)^T$. We compare the smoothed estimator with three alternative estimators. The first is the nonsmoothed estimator in (2.5), which was computed using the genetic algorithm, using the "genoud" function in R package "rgenoud" (Mebane, Jr. and Sekhon, 2011), as suggested in Zhang et al. (2012). The second is the estimator based on the hinge loss (Zhao et al., 2012), calculated using the function *owl* in the R package DTRlearn2 (Chen et al., 2019). The third is the estimator using logistic loss, calculated using the function *glmnet* in the R package glmnet (Friedman et al., 2010). Table 2.1 reports the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value.

The results in Table 2.1 demonstrates that the smoothed robust estimate has smaller bias and substantially smaller standard deviation comparing with the other three estimators, particular for the smaller sample size setting. It also leads to higher match ratio. Estimators using hinge loss and logistic loss are even not consistent when the sample size increases. For $n = 300$, we observe that in one or two of the 100 simulation runs the non-smooth estimator converges to the negative of the true value of β_1^{opt} (i.e., the algorithm converges to 1 when the true value is -1), which causes the non-zero variance. This is probably due to the fact nonsmooth estimation is less stable when the sample size is relatively small. In addition, the expected value functions with the true parameter $\boldsymbol{\beta}^{opt}$ and random policy are simulated via Monte Carlo simulation with 10^7 replicates; for Setting 1, the optimal value turns out to be 1.14, and the value function with random policy is -0.47; and for Setting 2, the true optimal value is 0.93, and the value function with random policy is -0.29. When taking the computation time into consideration, the nonsmoothed estimator requires about 4 seconds for each run, while the smoothed estimator only needs 0.002 seconds. This suggests a substantial reduction in computational costs.

Table 2.1: Monte Carlo estimates of the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value.

n	Method	β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	Match Ratio	$V_n(\widehat{\boldsymbol{\beta}}_n)$
				Setting 1			
300	Smooth	-0.05 (0.30)	0 (0)	0.01 (0.27)	0.04 (0.31)	99.35%	-0.02 (0.17)
	Nonsmooth	-0.29 (1.45)	0.00 (0.09)	0.12 (1.21)	0.24 (1.43)	96.67%	0.06 (0.17)
	Hinge	-0.46 (0.41)	0 (0)	0.04 (0.27)	-0.04 (0.29)	91.85%	-0.05 (0.18)
	Logistic	-0.46 (0.47)	0 (0)	0.06 (0.42)	0.26 (0.57)	94.17%	-0.02 (0.18)
500	Smooth	-0.01 (0.19)	0 (0)	0.01 (0.20)	0.02 (0.22)	99.73%	0.00 (0.13)
	Nonsmooth	-0.15 (0.41)	0 (0)	0.06 (0.36)	0.13 (0.42)	98.19%	0.05 (0.13)
	Hinge	-0.37 (0.30)	0 (0)	0.01 (0.18)	-0.06 (0.20)	92.93%	-0.03 (0.13)
	Logistic	-0.41 (0.29)	0 (0)	0.04 (0.30)	0.23 (0.36)	94.61%	-0.01 (0.13)
1000	Smooth	-0.01 (0.14)	0 (0)	0.00 (0.13)	0.01 (0.15)	99.88%	-0.01 (0.09)
	Nonsmooth	-0.07 (0.24)	0 (0)	0.02 (0.22)	0.06 (0.25)	99.04%	0.03 (0.09)
	Hinge	-0.36 (0.24)	0 (0)	0.01 (0.13)	-0.07 (0.14)	92.95%	-0.04 (0.09)
	Logistic	-0.38 (0.19)	0 (0)	0.02 (0.19)	0.18 (0.23)	94.61%	-0.02 (0.09)
				Setting 2			
300	Smooth	0.04 (0.26)	0 (0)	0.02 (0.24)	0.02 (0.18)	99.35%	-0.01 (0.15)
	Nonsmooth	-0.26 (0.76)	0.00 (0.06)	0.11 (0.71)	0.11 (0.37)	95.78%	0.07 (0.15)
	Hinge	-3.33 (79.42)	0 (0)	0.01 (0.22)	-0.09 (0.16)	76.19%	-0.06 (0.16)
	Logistic	-0.67 (5.13)	0.00 (0.06)	0.18 (3.33)	0.23 (2.96)	90.20%	-0.02 (0.16)
500	Smooth	0.02 (0.19)	0 (0)	0.02 (0.18)	0.00 (0.13)	99.65%	-0.01 (0.11)
	Nonsmooth	-0.16 (0.52)	0 (0)	0.06 (0.42)	0.06 (0.24)	97.37%	0.05 (0.11)
	Hinge	-0.64 (1.11)	0 (0)	0.02 (0.16)	-0.10 (0.12)	88.59%	-0.07 (0.12)
	Logistic	-0.43 (0.29)	0 (0)	0.03 (0.30)	0.12 (0.20)	92.08%	-0.03 (0.12)
1000	Smooth	-0.01 (0.14)	0 (0)	0.01 (0.13)	0.00 (0.09)	99.79%	-0.01 (0.08)
	Nonsmooth	-0.08 (0.21)	0 (0)	0.03 (0.22)	0.04 (0.17)	98.55%	0.03 (0.08)
	Hinge	-0.56 (0.24)	0 (0)	0.01 (0.12)	-0.10 (0.08)	89.69%	-0.06 (0.09)
	Logistic	-0.43 (0.20)	0 (0)	0.03 (0.20)	0.11 (0.15)	92.13%	-0.03 (0.09)

We next investigate the bootstrap confidence interval in Section 2.2.3. We construct 95% bootstrap confidence intervals for the parameters indexing the optimal treatment regime. Table 2.2 summarizes the empirical coverage probabilities and average interval lengths. We observe that the coverage probabilities are above 92.2% for sample sizes 500 and 1000, and above 91% for sample size 300. Despite the slight under coverage, the lengths of the confidence intervals are reasonable. As sample size increases, the length of the confidence

interval decreases significantly. Accurate finite-sample coverage is harder to achieve due to the model-free, nonparametric nature of our approach. See similar observations in simulations focusing on non-regularity settings for dynamic treatment regimes, for instance, Laber et al. (2014) and Chakraborty et al. (2013). As for computation time, on average one bootstrap run takes less than 0.2 seconds.

Table 2.2: Empirical coverage probabilities and average interval lengths of the 95% bootstrap confidence intervals for $\boldsymbol{\beta}^{opt}$

n		β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}
		Setting 1			
300	Coverage Rate	92.6%	100%	93.2%	91.0%
	Average Length	1.36	0	1.26	1.38
500	Coverage Rate	92.2%	100%	93.0%	92.6%
	Average Length	0.81	0	0.79	0.84
1000	Coverage Rate	92.6%	100%	94.0%	93.4%
	Average Length	0.54	0	0.53	0.56
		Setting 2			
300	Coverage Rate	93.4%	100%	92.6%	95.8%
	Average Length	1.12	0	1.01	0.71
500	Coverage Rate	94.2%	100%	93.8%	94.6%
	Average Length	0.75	0	0.72	0.51
1000	Coverage Rate	94.0%	100%	93.0%	95.4%
	Average Length	0.50	0	0.48	0.35

Finally, we explore several nonregular settings, where the optimal treatment regimes may be nonunique, motivated by Laber et al. (2014). In these cases, the parameter indexing the optimal treatment regime is not uniquely identifiable but inference for the optimal value may still be feasible. We focus here on the bootstrap confidence interval for the optimal value. In setting 3, the same data generative model as before is used with $\boldsymbol{\beta} = (1, 2, 0.02, 0)^T$. For setting 4 and 5, $\boldsymbol{\beta} = (-1, 1, 0, 0)^T$, however, the first random covariate x_1 is generated from the discrete uniform distribution on the set $\{-1, 0, 1, 2\}$ and

$\{1, 2\}$, respectively, instead of the standard normal distribution. For completeness, the bootstrap confidence intervals for the optimal value in setting 1 and setting 2 are also studied.

Let p denote the probability of generating a covariate vector $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{\beta} = 0$. This is a useful measure of the nonregularity of the model (Laber et al., 2014). According to this measurement, setting 1 – 3 are regular (R) cases with $p = 0$; while setting 4 and 5 are nonregular (NR) with $p = 0.25$ for setting 4 and $p = 0.5$ for setting 5.

Table 2.3 summarizes the empirical coverage rate and average length for the 95% bootstrap confidence intervals for the optimal value functions. The results demonstrate that the bootstrap confidence intervals for the optimal value have desirable coverage rates with reasonable interval lengths, even in the nonregular cases. For comparison, we also report the percentage of times these bootstrap confidence would cover the value function from a random policy. The percentage is really low, which implies that the proposed method performs much better than random assignment even in the nonregular cases.

Table 2.3: Empirical coverage probabilities and average interval lengths of the 95% confidence intervals for $V(\boldsymbol{\beta}^{opt})$

	Setting	1	2	3	4	5
n	Type	R	R	R	NR	NR
	Coverage Rate	93.0%	92.6%	96.4%	97.2%	95.4%
300	Average Length	0.67	0.61	0.78	0.40	0.41
	CR for random policy	0%	0%	0%	0%	31.2%
	Coverage Rate	93.8%	94.0%	96.0%	95.2%	94.4%
500	Average Length	0.52	0.47	0.62	0.31	0.31
	CR for random policy	0%	0%	0%	0%	12.4%
	Coverage Rate	93.6%	95.4%	97.0%	96.0%	96.0%
1000	Average Length	0.37	0.33	0.43	0.22	0.22
	CR for random policy	0%	0%	0%	0%	0.8%

2.5 A Real Data Example

We analyze a clinical data set from the Childhood Adenotonsillectomy Trial (CHAT). This is a randomized study designed to test whether early adenotonsillectomy (eAT, denoted as treatment 1) is helpful to improve neurocognitive functioning, behavior and quality of life for children with mild to moderate obstructive sleep apnea, compared with watchful waiting plus supportive care (WWSC, denoted as treatment 0), see Marcus et al. (2013). In this trial, 464 children with mild to moderate obstructive sleep apnea syndrome, ages 5 to 9.9 years, were randomly assigned to eAT and WWSC. Some biochemical and neurocognitive test results were recorded before the treatment and seven months after the treatment.

We consider the baseline Apnea-Hypopnea Index (AHI), with a natural log-transformation as recommended by Marcus et al. (2013), as an explanatory variable. AHI is the number of apneas or hypopneas recorded during the study per hour of sleep. It is an important measurement of the quality of sleep and is commonly used by doctors to classify the severity of sleep apnea. Marcus et al. (2013) suggested that black children tend to experience different improvements with eAT comparing with children from other races. We hence include race (binary, 1=African American, 0 for others) as another covariate. For the outcome variable, to balance the benefits and adverse effects from eAT, we adopt a composite score. The composite score uses the ratio of the follow-up AHI and baseline AHI (both with natural log-transformations) as an effective measure of benefit. On the other hand, it takes into account the adverse events documented according to the CHAT study manual of procedures as penalty.

We estimate the optimal treatment regime in the class of treatment regimes $\mathbb{D} = \{I(\beta_0 + \beta_1 \text{AHI} + \beta_2 \text{race} > 0) : |\beta_1| = 1\}$. The kernel function $K(\cdot)$ and the bandwidth selection are the same as in Section 2.4. The smoothed estimator for the baseline AHI is normalized to 1, the race is 0.56, with $(0.34, 0.97)$ as the 95% bootstrap confidence interval, and the intercept is 0.39, with confidence interval $(0.22, 0.65)$. The confidence intervals suggest that the

coefficients are all significantly different from 0. The analysis suggests that it is reasonable to assign WWSC to those children with milder symptoms (lower AHI). It also suggests that black children display more improvement in the AHI scale with eAT. The results are consistent with those observed empirically in Redline et al. (2011), Marcus et al. (2013) and Dean et al. (2016). The average outcome with randomized treatment is 0.288. The estimated average outcome corresponding to the estimated optimal treatment regime is 0.063, with a 95% bootstrap confidence interval $(-0.126, 0.260)$. This suggested a significant reduction of the composite outcome score when applying the optimal treatment regime. To compare with the smoothed estimator, we also calculate the nonsmoothed estimator, whose coefficients are 1 for baseline AHI, -0.19 for the race, and -0.40 for the intercept. Its estimated optimal value is -0.034. The nonsmoothed estimators are significantly different from the smoothed ones. In Example 4 of Appendix A.10 in the supplementary, we demonstrate based on five-fold cross-validation that for this real data example, the nonsmoothed estimator is quite unstable.

2.6 Discussions

2.6.1 Extension to other settings

The method we propose can be extended to observational studies using the inverse probability weighting approach. Assume the propensity score $\pi(x) = P(A = 1|\boldsymbol{x})$ can be modeled as $\pi(x, \xi)$ where ξ is a finite-dimensional parameter (e.g., via logistic regression). Let $\widehat{\xi}$ be an estimate of ξ. Under the commonly adopted assumption of no unmeasured confounding, a smoothed robust estimator for $\boldsymbol{\beta}_0$ can be constructed as

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1} \sum_{i=1}^{n} \frac{\big[A_i K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) + (1-A_i)\big\{1 - K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\big\}\big]Y_i}{A_i\pi(x,\widehat{\xi}) + (1-A_i)(1-\pi(x,\widehat{\xi}))}. \tag{2.13}$$

Example 3 in Appendix A.10 of the supplementary material confirms that this smoothed estimator provides accurate estimation for the optimal treatment regime when the propensity score model is correctly specified. The estimator in (2.13) can also be extended to be doubly robust similarly as in Zhang et al. (2012). Due to the presence of nuisance parameter, the theory of asymptotic normality and inference is more technically involved. This will be a future research topic.

It is worth pointing out that our method is applicable to binary response, as binary random variable is sub-Gaussian after centering. Example 1 in Appendix A.10 of the supplementary materiel demonstrates that our estimation and inference procedures work effectively for binary responses. For survival outcome under random censoring, our method can be extended to obtain a robust procedure for estimating the optimal treatment regime maximizing the restricted mean survival time, similarly as in Zhao et al. (2015b). Let $\widetilde{T}$ denote the survival time. Let $T = \min\{\widetilde{T}, \tau\}$ be the outcome of interest, where τ is the time till the end of the study. Let C denote the censoring time and $\Delta = I(T < C)$ be the censoring indicator. We observe $Y = \min\{T, C\}$. Based on the observed data $\{Y_i, \boldsymbol{x}_i, \Delta_i, A_i\}$, $i = 1, \ldots, n$ from a randomized trial, the smoothed estimator can be constructed as

$$\arg\max_{\boldsymbol{\beta}\in\mathbb{B}} \frac{2}{n}\sum_{i=1}^{n} \frac{\big[A_i K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) + (1-A_i)\big\{1 - K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\big\}\big]}{\widehat{G}_C(Y_i|\boldsymbol{x}, A)}\Delta_i Y_i,$$

where $G_C(t|\boldsymbol{x}, A) = P(C > t|\boldsymbol{x}, A)$ is the conditional survival function of the censoring time C given $(\boldsymbol{X}, A)$, and $\widehat{G}_C(\cdot|\boldsymbol{x}, A)$ is an estimator of $G_C(\cdot|\boldsymbol{x}, A)$.

2.6.2 On the identifiability condition

The asymptotic normality results can be established under alternative identifiability constraint such as the requirement that the L_1/L_2 norm of $\boldsymbol{\beta}$ is 1, or identifiability of $\boldsymbol{\beta}$ up to a scale. However, this usually leads to more technically involved proof as $\boldsymbol{\beta}$ is constrained to be the boundary point of a unit sphere and $V(\boldsymbol{\beta})$ does not have a derivative at $\boldsymbol{\beta}$. This issue

was often ignored in the theory development in many existing literature, which only adjust for the constraint in an ad-hoc way in the numerical implementation. See Zhu and Xue (2006) for more discussions in an index model setting and a careful delete-one-component method to handle this rigorously.

For identifiability, we assume that there exists a covariate whose conditional distribution given the other covariates is absolutely continuous. This is a common assumption for index model and is satisfied in many real applications. In practice, domain experts may help suggest such a candidate continuous covariate and the statisticians can run confirmatory analysis (e.g., comparing the conditional treatment effect conditional on this covariate) to verify if this is a viable choice. In the case when all relevant covarites are discrete (e.g, gender, race), the problem reduces to comparing a finite number of decision rules and the main target of inference is arguably the optimal value. Our simulation settings 4 & 5 only include discrete variables in the optimal regime. The simulation results in Table 2.3 show that our proposed bootstrap confidence interval still provides reasonable empirical coverage probability for the optimal value in discrete cases.

2.6.3 Non-regular settings

The optimal treatment regime may not be unique if there exists a subpopulation who responds similarly to the two treatment options. In such a setting, the complexity of nonregularity arises, see the discussions in Robins (2004), Moodie and Richardson (2010), Laber et al. (2014), Song et al. (2015), and Luedtke and Van Der Laan (2016). Uniform inference under nonregularity or exceptional laws is a challenging problem.

Although our theory does not apply to this scenario, our simulation results show that our bootstrap confidence interval for the optimal value function displays a fair degree of robustness in the two examples where nonregularity occurs. As an example, in simulation setting 5, if $x_1 = 1$, then the subject responds the same to the two treatment options; while

if $x_1 = 2$, the subject benefits from treatment 1. There are four decision rules of interest for this example. The optimal treatment rule is nonunique as one may assign either treatment 0 (say no treatment or a standard, less expensive treatment) or treatment 1 to those subjects with $x_1 = 1$. A relative simple approach to breaking the nonuniqueness is to introduce a secondary criterion. For example, one may argue that under the principle of avoiding over-treatment, there exists a unique optimal decision rule of interest, in this case $I(x_1 = 2)$, which would not assign treatment 1 when ambiguity exists in order to reduce costs and avoid potential risks. Based on the sample, this unique optimal treatment regime can be consistently estimated by selecting the decision rule that maximizes the sample average treatment effect while treating the smallest proportion of the population.

There are additional inference targets that have rarely been discussed in the literature, that is, inference about the linear combination in the rule $\boldsymbol{x}^T\boldsymbol{\beta}$ or about the rule itself $I(\boldsymbol{x}^T\boldsymbol{\beta} > 0)$. These two quantities are of interest in clinical practice as they indicate how much confidence we can put on the prescribed optimal decision. We are currently studying these inference problems and will report the results in a future article.

Chapter 3

Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

3.1 Introduction

Precision medicine is an innovative practice for disease treatment that takes into account individual variability in genes, environment, and lifestyle for each patient. Substantial efforts have recently been devoted to studying how to estimate the optimal personalized treatment regime given the individual-level information, which aims to yield the best expected outcome if the treatment regime is followed by each individual in the population. Several successful approaches have been developed for this estimation problem, including Q-learning and A-learning based methods (Watkins and Dayan, 1992; Robins et al., 2000; Murphy, 2003; Moodie and Richardson, 2010; Qian and Murphy, 2011), and classification-based methods (Zhang et al., 2012; Zhao et al., 2012, 2015a; Wang et al., 2018; Qi et al., 2018), among others. We refer to Chakraborty and Moodie (2013) and Kosorok and Moodie (2016) for a general introduction to this area and other relevant references.

Inference or uncertainty quantification is important in practice. This chapter studies the following inference problem for optimal personalized decision making: suppose we have

a large number of covariates (e.g., hundreds of genes), how will we determine if a given subset of covariates (e.g., genes associated with a given biological pathway) is relevant for making the optimal treatment recommendation? Scientifically, this knowledge would enable the doctors and researchers to identify critical characteristics (e.g., gender, age, gene pathways) that are influential for the optimal decision. It also helps gain insight into what information is worth collecting to be more cost effective, given the possibility of measuring a large number of variables (genetic, clinic, etc).

In the last a few years, important progress has been made in inference with optimal decision rules. Laber et al. (2014) developed a novel locally consistent adaptive confidence interval for the Q-learning approach. Chakraborty et al. (2013) proposed a practically convenient adaptive m-out-of-n bootstrap method for inference for Q-learning. Song et al. (2015) studied penalized Q-learning. Jeng et al. (2018) developed Lasso-based debiased procedure for A-learning. Different but related, Chakraborty et al. (2014) and Luedtke and Van Der Laan (2016), Zhu et al. (2018) developed confidence intervals for another quantity of interest: the value function. However, existing work mostly deal with the classical asymptotic setting of fixed p and large n, where p is the number of covariates and n is the sample size, and have not addressed the challenge of inference with high-dimensional variables. Moreover, the aforementioned work often assumes that the interaction between the covariates and the treatment has a known functional form.

Motivated by the overarching goal of precision medicine to incorporate genetic information (e.g, measurements on thousands of genes) in the decision making process, this chapter investigates inference about the effect of a group of variables on the optimal decision rule in the high-dimensional setting. We adopt a semiparametric model-assisted approach for optimal decision estimation and inference. The semiparametric structure permits nonlinear interaction effect between the covariates and treatment via an unknown smooth link function. This semiparametric framework incorporates many existing models as special cases.

Our results make a significant contribution to inference for optimal personalized decision. We derive simultaneous confidence intervals for inference on a group of variables while allowing the number of covariates to exceed the sample size. The confidence intervals enjoy the *honest* property in the following sense

$$\sup_{\boldsymbol{\beta}_0:||\boldsymbol{\beta}_0||_0\leqslant s}\sup_{\alpha\in(0,1)}\left|P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j-\beta_{0j}|\leqslant c^*_{1-\alpha}\Big)-(1-\alpha)\right|=o(1),$$

where $\boldsymbol{\beta}_0 = (\beta_{01},\ldots,\beta_{0p})^T$ is the population parameter indexing the optimal treatment regime, $\widetilde{\beta}_j$'s denote debiased estimators that will be introduced later, $\mathcal{G}$ denotes the group of variables of interest, $||\cdot||_0$ denotes the l_0 norm of a vector, and s is a positive integer denoting the sparsity size. The significance of the honest property is that the coverage probability is asymptotically valid uniformly over a class of s-sparse models. An immediate implication is that it relaxes the assumption on signal strength and does not require the zero and nonzero effects to be well-separated (so-called $\beta_{\min}$ condition). In particular, this procedure does not require the initial estimator to achieve perfect model selection. It avoids the problems associated with the nonuniformity of the limiting theory for penalized estimators, see discussions in Li et al. (1989); Pötscher (2009); Van de Geer et al. (2014); McKeague and Qian (2015), among others. It is also worth noting that the number of variables in $\mathcal{G}$ can be either small or large. For example, one may be interested in assessing how a group of genes corresponding to a particular biological pathway, the size of which can be comparable with or even larger than the sample size, affect optimal decision making. The critical value $c^*_{1-\alpha}$ is obtained using a wild bootstrap procedure, which automatically accounts for the dependence of the coordinates for testing component-wise hypotheses and leads to more accurate finite-sample performance.

This research also makes new contributions to statistical inference in high-dimensional semiparametric models. When the interaction effects are nonlinear, the estimated parameter indexing the optimal decision rule does not necessarily correspond to the solution of a

convex problem. Moreover, the estimated nonparametric component in the profiled score function creates substantial barriers for estimation, computation and inference in high dimension. Recently, confidence intervals and hypothesis testing have been thoroughly investigated for high-dimensional linear regression and generalized linear regression, see Belloni et al. (2014); Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014), and much of the subsequent work in this area. However, the existing tools are insufficient to address the challenges in our setting for inference for optimal decision making in a flexible semiparametric framework. Adopting tools from modern empirical process and random matrix theory, we establish that a local restricted strong convexity condition holds with high probability in high dimension and that any local sparse solution of the estimation equation can achieve desirable estimation accuracy. We further verify that valid honest uniform inference can be obtained based on a debiased version of a local solution. Moreover, we propose a new algorithm for efficient computation in high dimension.

The remainder of this chapter is organized as follows. Section 3.2 introduces the new methodology. Section 3.3 studies the statistical properties. Section 3.4 provides the details on computation and reports numerical results from Monte Carlo studies. Section 3.5 illustrates the new methods on a real data example from a diabetes study. Section 3.6 discusses some extensions. All the regularity conditions, several useful technical lemmas and the proofs are given in Appendix B.

3.2 Methodology

3.2.1 A Semiparametric Framework

For notational simplicity, we will focus on the binary decision setting. Let $A \in \mathcal{A} = \{0, 1\}$ denote a binary treatment and $\boldsymbol{x} \in \mathcal{X}$ denote a p-dimensional vector of baseline covariates. Let Y denote the outcome of interest. Without loss of generality, we assume a larger value

of the outcome is preferred. The observed data consist of $\{(\boldsymbol{x}_i, A_i, Y_i) : i = 1, \cdots, n\}$. We are interested in the setting where $p \gg n$.

A treatment regime is an individualized decision rule that can be represented as a function $d(\boldsymbol{x}) : \mathcal{X} \to \mathcal{A}$. The optimal treatment regime is defined as the decision rule which, if followed by the whole population, will achieve the largest average outcome. Formally, it is defined using the potential outcome framework in causal inference (Neyman, 1990; Rubin, 1978). Let $Y^*(a)$ be the potential outcome had the subject been assigned to treatment $a \in \{0, 1\}$. Given a treatment regime $d(\boldsymbol{x})$, the corresponding potential outcome is $Y^*(d) = Y^*(1)d(\boldsymbol{x}) + Y^*(0)(1 - d(\boldsymbol{x}))$. The optimal treatment regime is defined as $d^{\text{opt}}(\boldsymbol{x}) = \arg\max_d \mathrm{E}\{Y^*(d)\}$. It is now well known that $d^{\text{opt}}(\boldsymbol{x}) = \arg\max_{a \in \mathcal{A}} \mathrm{E}(Y|\boldsymbol{x}, A = a)$ (Qian and Murphy, 2011).

This chapter considers a flexible semiparametric framework for optimal treatment regime estimation and inference in the high-dimensional setting. Specifically, we assume

$$Y_i = g(\boldsymbol{x}_i) + (A_i - 1/2) f_0(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) + \epsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where ϵ_i is the random error, $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \cdots, \beta_{0p})^T$, $g(\boldsymbol{x}_i)$ is the unknown main effect, and $f_0(\cdot)$ is an unknown increasing function that describes the interaction between the treatment and covariates. It is assumed that $\mathrm{E}(\epsilon_i|\boldsymbol{x}_i) = 0$, $i = 1, \ldots, n$. For identifibility, we assume $f_0(0) = 0$ and $|\beta_{01}| = 1$. Under model (3.1), the optimal treatment regime is $d^{\text{opt}}(\boldsymbol{x}) = \mathrm{I}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 > 0)$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Note that the class of index rules are popular in practice due to its interpretability.

Existing work on inference for optimal treatment regime is mostly based on a parametric generative model, which is prone to model misspecification. The semiparametric structure alleviates this difficulty. In particular, it allows for possible nonlinear interaction effects between the covariates and treatment. It also circumvents the curse of dimensionality associated with a fully nonparametric model.

Our goal is to estimate $\boldsymbol{\beta}_0$ and make inference on its components in the high-dimensional setting. In the special case $f_0(u) = u$, which is popularly used in practice, the problem can be formulated as a high-dimensional convex estimation problem. However, when f_0 is nonlinear, it generally leads to a high-dimensional nonconvex problem. Both estimation and inference need to overcome new challenges.

3.2.2 Profiled Semiparametric Estimation

We start with introducing a penalized profiled semiparametric estimation equation for estimating the parameter indexing the optimal treatment regime. We consider data from a random experiment, that is, $P(A_i = 0) = P(A_i = 1) = 1/2$, $i = 1, \ldots, n$. Extension to data from observational studies is discussed in Section 3.6. Inspired by an observation made for the linear model (Tian et al., 2014), we observe

$$2(2A_i - 1)Y_i = f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) + 2(2A_i - 1)\big[\epsilon_i + g(\boldsymbol{x}_i)\big]. \tag{3.2}$$

Let $\widetilde{Y}_i = 2(2A_i - 1)Y_i$ be the modified response, and let $\widetilde{\epsilon}_i = 2(2A_i - 1)\big[\epsilon_i + g(\boldsymbol{x}_i)\big]$ be the modified error. We have

$$\mathrm{E}\{\widetilde{Y}_i|\boldsymbol{x}_i\} = f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0). \tag{3.3}$$

In the ideal situation where the link function f_0 is known, we have $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathrm{E}\big[\widetilde{Y}_i - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^2$. It is noteworthy that for a nonlinear function f_0, the objective function is usually nonconvex in $\boldsymbol{\beta}$. Ichimura (1993) carefully studied the properties of the global minimizer for a semiparametric nonlinear least-squares approach in the classical finite-dimensional setting.

To estimate $\boldsymbol{\beta}_0$ in the high-dimensional setting with an known f_0, we consider a penalized profiled semiparametric estimation equation. To motivate, we observe that in the

ideal situation where f_0 is known a priori, $\boldsymbol{\beta}_0$ satisfies the following unbiased estimating equation

$$\mathrm{E}\big\{\big[\widetilde{Y}_i - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i\big\} = \mathbf{0}, \tag{3.4}$$

where $f_0'(\cdot)$ denotes the derivative of $f_0(\cdot)$. We will replace the unknown f_0 and f_0' by their respective profiled nonparametric estimator, and consider an appropriately penalized version of the estimated score function to handle the high-dimensional covariates.

We summarize the main steps of estimation as follows. Define $G(t|\boldsymbol{\beta}) = \mathrm{E}\{\widetilde{Y}|\boldsymbol{x}^T\boldsymbol{\beta} = t\}$. Note that $G(t|\boldsymbol{\beta}_0) = f_0(t)$. However, when $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $G(t|\boldsymbol{\beta})$ usually has a functional form different from f_0. Consider the Nadaraya-Watson kernel estimator for $G(t|\boldsymbol{\beta})$:

$$\widehat{G}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t,\boldsymbol{\beta})\widetilde{Y}_i, \tag{3.5}$$

where $K_h(z) = h^{-1}K(z/h)$, and $W_{ni}(t,\boldsymbol{\beta}) = \frac{K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{j=1}^{n}K_h(t-\boldsymbol{x}_j^T\boldsymbol{\beta})}$. Write $G^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt}G(t|\boldsymbol{\beta})$ and $W_{ni}^{(1)}(t,\boldsymbol{\beta}) = \frac{d}{dt}W_{ni}(t,\boldsymbol{\beta})$. Then the kernel estimator for the derivative $G^{(1)}(t|\boldsymbol{\beta})$ is

$$\widehat{G}^{(1)}(t|\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}^{(1)}(t,\boldsymbol{\beta})\widetilde{Y}_i. \tag{3.6}$$

Write $G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta}) = \mathrm{E}\{\widetilde{Y}|\boldsymbol{x}^T\boldsymbol{\beta}\}$. Denote $G(\boldsymbol{x}^T\boldsymbol{\beta}) = G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$, $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}) = G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$, $\widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}) = \widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $\widehat{G}^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}) = \widehat{G}^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ for notational simplicity. To estimate $\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta})$ and $\widehat{G}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta})$, we employ the following leave-one-out estimators

$$\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta}) = \sum_{i\neq j} W_{nij}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta})\widetilde{Y}_i, \tag{3.7}$$

$$\widehat{G}(\boldsymbol{x}_j^T\boldsymbol{\beta}) = \sum_{i\neq j} W_{nij}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta})\widetilde{Y}_i, \tag{3.8}$$

where $W_{nij}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{K_h(\boldsymbol{x}_j^T\boldsymbol{\beta}-\boldsymbol{x}_i^T\boldsymbol{\beta})}{\sum_{k\neq j}K_h(\boldsymbol{x}_j^T\boldsymbol{\beta}-\boldsymbol{x}_k^T\boldsymbol{\beta})}$, and $W_{nij}^{(1)}(\boldsymbol{x}_j^T\boldsymbol{\beta},\boldsymbol{\beta}) = \frac{d}{dt}W_{nij}(t,\boldsymbol{\beta})\Big|_{t=\boldsymbol{x}_j^T\boldsymbol{\beta}}$. The estimated semiparametric estimating function motivated by (3.4) is

$$\boldsymbol{S}_n(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^{n}\big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i. \tag{3.9}$$

In the high-dimensional setting, the estimating equation $\boldsymbol{S}_n(\boldsymbol{\beta}) = \mathbf{0}$ is ill-posed when $p \gg n$. A solution $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1,\ldots,\widehat{\beta}_p)^T$ is defined as one that satisfies the following penalized semiparametric profiled estimating equation

$$\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\boldsymbol{\kappa}} = \mathbf{0}, \tag{3.10}$$

where $\lambda > 0$ is a tuning parameter, $\widehat{\boldsymbol{\kappa}} = (\widehat{\kappa}_1,\ldots,\widehat{\kappa}_p)^T \in \partial||\widehat{\boldsymbol{\beta}}||_1$ with $||\widehat{\boldsymbol{\beta}}||_1$ denoting the l_1 norm of the vector $\widehat{\boldsymbol{\beta}}$ and $\partial||\widehat{\boldsymbol{\beta}}||_1$ denoting the subdifferential of $||\widehat{\boldsymbol{\beta}}||_1$, that is $|\widehat{\kappa}_j| \leqslant 1$, $\forall\ j$ and $\widehat{\kappa}_j = \text{sign}(\widehat{\beta}_j)$, if $\widehat{\beta}_j \neq 0$. Note that the estimating equation may have multiple solutions. The theory we develop in Section 3.3.1 provides a near-optimal error bound for any local solution of the estimating equation. The satisfactory performance of the proposed profiled estimator is demonstrated in the numerical simulations in Section 3.4.3.

3.2.3 Inference on the Optimal Decision Rule

We now develop inference procedures to quantify the importance of the covariates on optimal decision making. We will first study confidence intervals for the individual components of $\boldsymbol{\beta}_0$ via debiasing a local solution to the semiparametric estimating equation (3.10). This generalizes the work of debiased confidence intervals for high-dimensional linear regression in Zhang and Zhang (2014) and Van de Geer et al. (2014) to the semiparametric setting where the initial estimator is an estimating equation solution and an an estimated infinite-dimensional functional is present. The theory for semiparametric inference in high dimension is highly nontrivial and is carefully studied in Section 3.3. We further investigate

a wild bootstrap procedure for testing a general group hypothesis, which aims to achieve accurate finite-sample performance.

Let $\widehat{\boldsymbol{\beta}}$ denote a solution satisfying (3.10). In the high-dimensional linear regression setting, the main idea of debiased estimators is to invert the Karush–Kuhn–Tucker (KKT) condition of the lasso. Inspired by this idea, we consider the following debiased estimator

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}^T \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}), \tag{3.11}$$

where $\widehat{\boldsymbol{\Theta}}$ is an approximation to the inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}})$, the gradient of $\boldsymbol{S}_n(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. To construct the approximate inverse $\widehat{\boldsymbol{\Theta}}$, we propose a nodewise Dantzig estimator. Specifically, given $\boldsymbol{\beta} \in \mathbb{R}^p$ and a positive number η, for $j = 1, \cdots, p$, define

$$\boldsymbol{d}_j(\boldsymbol{\beta}, \eta) = \min_{\boldsymbol{v} \in \mathbb{R}^{p-1}} ||\boldsymbol{v}||_1 \quad \text{subject to } \left|\left| n^{-1} \sum_{i=1}^{n} [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta})]^2 (x_{i,j} - \boldsymbol{x}_{i,-j}^T \boldsymbol{v}) \boldsymbol{x}_{i,-j} \right|\right|_\infty \leqslant \eta, \tag{3.12}$$

where $||\cdot||_\infty$ denotes the infinity norm of a vector, $x_{i,j}$ denotes the j^{th} entry of the vector $\boldsymbol{x}_i$, and $\boldsymbol{x}_{i.-j}$ denotes the $(p-1)$-subvector of $\boldsymbol{x}_i$ that excludes the j^{th} entry. Furthermore, for $j = 1, \cdots, p$, we define

$$\begin{aligned}
\boldsymbol{\phi}_j(\boldsymbol{\beta}, \eta) &= \Big(- \big(\boldsymbol{d}_j(\boldsymbol{\beta}, \eta)\big)_{1:(j-1)}^T, 1, - \big(\boldsymbol{d}_j(\boldsymbol{\beta}, \eta)\big)_{j:(p-1)}^T \Big)^T, && (3.13)\\
\tau_j^2(\boldsymbol{\beta}, \eta) &= n^{-1} \sum_{i=1}^{n} [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta})]^2 x_{i,j} \boldsymbol{x}_i^T \boldsymbol{\phi}_j(\boldsymbol{\beta}, \eta), && (3.14)\\
\boldsymbol{\theta}_j(\boldsymbol{\beta}, \eta) &= \tau_j^{-2}(\boldsymbol{\beta}, \eta) \boldsymbol{\phi}_j(\boldsymbol{\beta}, \eta), && (3.15)
\end{aligned}$$

where for a vector $\boldsymbol{w} = (w_1, \cdots, w_p)^T$, $(\boldsymbol{w})_{i:j}$ returns the subvector $(w_i, \cdots, w_j)^T$ for $1 \leqslant i \leqslant j \leqslant p$; and returns an empty vector otherwise. Given $\widehat{\boldsymbol{\beta}}$, denote $\widehat{\boldsymbol{d}}_j = \boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$, $\widehat{\tau}_j^2 = \tau_j^2(\widehat{\boldsymbol{\beta}}, \eta)$, and $\widehat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}}, \eta)$. The approximate inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}})$ is constructed as

$$\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}_1, ..., \widehat{\boldsymbol{\theta}}_p).$$

The validity of $\widehat{\boldsymbol{\Theta}}$ an approximation to the inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}})$ is given in Lemma 3.2 in Section 3.3.2.

Section 3.3 will present the asymptotic normality for the desparsified estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \cdots, \widetilde{\beta}_p)^T$, defined in (3.11) with the above approximate inverse $\widehat{\boldsymbol{\Theta}}$. In addition, the empirical estimator of the asymptotic covariance matrix of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) := \widehat{\boldsymbol{\Theta}}^T \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})\big]^2 [\widehat{G}^{(1)}(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})]^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \Big\} \widehat{\boldsymbol{\Theta}}. \tag{3.16}$$

Its consistency is shown in Corollary B.12 of the supplementary material. An asymptotic $100(1-\alpha)\%$ confidence interval for β_{0j}, $j = 2, \cdots, p$, is given by

$$\Big\{ \widetilde{\beta}_j - \Phi^{-1}(1-\alpha/2)\big(\widehat{\Sigma}_{jj}/n\big)^{1/2}, \widetilde{\beta}_j + \Phi^{-1}(1-\alpha/2)\big(\widehat{\Sigma}_{jj}/n\big)^{1/2} \Big\}, \tag{3.17}$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution, and $\widehat{\Sigma}_{jj}$ denotes the j^{th} diagonal entry of $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}})$. Corollary 3.1 in Section 3.3 justifies the asymptotic uniform validity of this marginal confidence interval.

Next, we consider the following more general simultaneous testing problem:

$$H_{0,\mathcal{G}} : \beta_{0j} = 0 \text{ for all } j \in \mathcal{G} \quad \text{versus} \quad H_{1,\mathcal{G}} : \beta_{0j} \neq 0 \text{ for some } j \in \mathcal{G}, \tag{3.18}$$

where $\mathcal{G}$ is a prespecified subset of variables, whose size may depend on the sample size n. Such a hypothesis naturally arises in the high-dimensional setting. For example, researchers may want to test whether a gene pathway, consisting of multiple genes (the number of genes in the pathway can be large relative to the sample size) for the same biological functions, is important for optimal treatment regime recommendation. For this purpose, we propose an effective bootstrap procedure. Although the asymptotic normal distribution of the debiased estimator (see Theorem 3.2) allows for construction of confidence intervals for individual coefficients or confidence regions for fixed-dimensional vector of coeffi-

cients. Applying it to make inference for groups of variables when the group size diverges (potentially larger than n) is not straightforward. Second, directly using the asymptotic distribution has been empirically observed to sometimes lead to undercoverage for confidence intervals for nonzero coefficients in finite samples. The bootstrap procedure we study below automatically accounts for the dependence structure of the variables in the group and provides more accurate finite-sample performance.

When deriving the asymptotic property of the debiased estimator (in the proof of Theorem 3.2), it is observed that the asymptotic distribution of $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is determined by $\sqrt{n}\widehat{\boldsymbol{\Theta}}^T \boldsymbol{S}_n(\boldsymbol{\beta}_0)$. This suggests that we can approximate the asymptotic distribution of $\sqrt{n}(\widetilde{\beta}_j - \beta_{0j})$ by the distribution of the following multiplier bootstrap statistic

$$\delta_j^* := \tfrac{1}{n}\textstyle\sum_{i=1}^n r_i \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j, \tag{3.19}$$

where $r_1, \cdots, r_n$ are i.i.d. standard normal random variables, independent of the data. Let $c_{1-\alpha}^*$ be the upper α-quantile of the distribution of $\max_{j\in\mathcal{G}} |\delta_j^*|$ conditional on the data, which can be easily simulated by generating multiple independent copies of the random weights. We reject the null hypothesis at level α if $\max_{j\in\mathcal{G}} |\widetilde{\beta}_j| > c_{1-\alpha}^*$. The asymptotic validity of the bootstrap procedure is formally established in Section 3.3. Its satisfactory performance is demonstrated in the numerical simulations in Section 3.4.3.

3.3 Statistical Properties

In this section, we carefully study the statistical theory for estimation and inference. The proof of the theory can be found in the supplementary material.

3.3.1 Theory for Estimation

To estimate and make inference about the optimal treatment regime, a significant challenge we face in the high-dimensional semiparametric framework is that the corresponding estimation problem is not necessarily convex. To address this challenge, we first establish in Lemma 3.1 that the estimated score function $\boldsymbol{S}_n(\cdot)$ in (3.9) possesses an important local restricted strong convexity property with high probability. Theorem 3.1 then shows that all local sparse solutions within a small ball of $\boldsymbol{\beta}_0$ enjoy a near-optimal error rate under mild conditions. In the sequel, we use $a \vee b$ to denote $\max(a, b)$, and $a \wedge b$ to denote $\min(a, b)$. Let $s = ||\boldsymbol{\beta}_0||_0$ denote the sparsity size of the $\boldsymbol{\beta}_0$, the population parameter indexing the optimal treatment regime.

Lemma 3.1 (Local restricted strong convexity property)
Assume conditions (B1)–(B5) in the appendix are satisfied. If $d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5} \leqslant h < 1$ for some constant $d_0 > 0$, then there exist universal positive constants c_0, c_1, c_2 and r, which do not depend on n, p and $\boldsymbol{\beta}_0$, such that

$$
\begin{aligned}
&P\Big(\big\langle \boldsymbol{S}_n(\boldsymbol{\beta}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0), \boldsymbol{\beta} - \boldsymbol{\beta}_0 \big\rangle \geqslant c_0||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2^2 - c_1 h^2 ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1,\ \forall\, \boldsymbol{\eta} \in \mathbb{B}\Big) \\
&\geqslant 1 - \exp\big\{ - c_2\big[(ns\log p)^{1/3} \wedge \log p\big]\big\}
\end{aligned}
$$

for all n sufficiently large, where $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 \leqslant r, ||\boldsymbol{\beta}||_0 \leqslant ks\}$ and $k > 1$ is a positive constant, □

Remark 3.1
Lemma 3.1 characterizes the local geometry of the profiled score function. For high-dimensional regression with convex loss function such as L_1 penalized high-dimensional linear regression, restricted strong convexity plays an important role on the performance of the regularized estimator (Negahban et al., 2012). Local restricted strong convexity condition were investigated in Loh and Wainwright (2015) and Mei et al. (2018) in their work

on parametric high-dimensional model with nonconvex loss functions. Those results do not apply to our setting as the score function involves an estimated infinite-dimensional parameter. □

Theorem 1 below presents non-asymptotic high probability error bounds for any local sparse solution $\widehat{\boldsymbol{\beta}}$ that satisfies the penalized profiled estimation equation (3.10).

Theorem 3.1

Assume conditions (B1)–(B5) in the appendix are satisfied. Suppose $\lambda = d_1 \max\{h^2, \sqrt{\frac{\log p}{n}}\}$ for some constant $d_1 > 0$, and $d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5} \leqslant h \leqslant d_0 n^{-1/6}$ for some constant $d_0 > 0$. Then there exist universal positive constants c_0 and c_1 such that for any solution $\widehat{\boldsymbol{\beta}}$ in $\mathbb{B}$, we have

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2 \leqslant \frac{6}{c_0}\lambda\sqrt{s}, \quad ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 \leqslant \frac{24}{c_0}\lambda s,$$

with probability at least $1 - \exp(-c_1 \log p)$, and all n sufficiently large. □

Remark 3.2

Theorem 3.1 shows that under some mild regularity conditions, local solutions of the profiled estimation equation (3.10) enjoy desirable estimation error rates, similarly as Lasso for high-dimensional linear regression. Carefully going through the proof of the theorem also reveals that the above error bounds hold uniformly for all $\boldsymbol{\beta}_0$ such that $||\boldsymbol{\beta}_0||_0 \leqslant s$. Lemmas B.4–B.5 in the appendix establish the uniform convergence rates for the nonparametric estimator $\widehat{G}(t|\boldsymbol{\beta})$ and $\widehat{G}^{(1)}(t|\boldsymbol{\beta})$, which are of independent interest. □

3.3.2 Theory for Inference

To study the theory for inference, we first define several population quantities. Denote the matrix $\boldsymbol{\Omega} = \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\big\}$, and its inverse $\boldsymbol{\Omega}^{-1} := \boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p)$. For

$j = 1, \ldots, p$, let $\boldsymbol{\Omega}_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the submatrix of $\boldsymbol{\Omega}$ with its j^{th} row and j^{th} column removed; similarly $\boldsymbol{\Omega}_{-j,j} \in \mathbb{R}^{p-1}$ denotes the j^{th} column of $\boldsymbol{\Omega}$ with its j^{th} entry removed. Assume $\boldsymbol{\Omega}_{-j,-j}$ is positive definite. Define $\boldsymbol{d}_{0j} = (\boldsymbol{\Omega}_{-j,-j})^{-1}\boldsymbol{\Omega}_{-j,j}$, $s_j = ||\boldsymbol{d}_{0j}||_0$, $\widetilde{s} = \max_{1\leqslant j\leqslant p} s_j$ and $\tau_{0j}^2 = \boldsymbol{\Omega}_{j,j} - \boldsymbol{d}_{0j}^T\boldsymbol{\Omega}_{-j,j} = (\Theta_{j,j})^{-1}$.

We first present below a useful lemma regarding the properties of the approximate inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}})$ in Section 3.2.3.

Lemma 3.2

Assume the conditions of Theorem 3.1 are satisfied. Let $\eta = d_2 h$ for some positive constant $d_2 > 0$. If $\eta\widetilde{s} \leqslant d_0$ and $d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5} \leqslant h \leqslant d_0 n^{-1/6}$ for some constant $d_0 > 0$, then there exist universal positive constants c_0 and c_1 such that results (1)-(4) below hold with probability at least $1 - \exp(-c_1 \log p)$ for all n sufficiently large:

(1) $\max_{1\leqslant j\leqslant p} s_j^{-1/2}||\widehat{\boldsymbol{d}}_j - \boldsymbol{d}_{0j}||_2 \leqslant \frac{8\eta}{a^2\xi_1}$, and $\max_{1\leqslant j\leqslant p} s_j^{-1}||\widehat{\boldsymbol{d}}_j - \boldsymbol{d}_{0j}||_1 \leqslant \frac{16\eta}{a^2\xi_1}$;

(2) $\max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^2 - \widehat{\tau}_j^2| \leqslant c_0\eta$, and $\max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^{-2} - \widehat{\tau}_j^{-2}| \leqslant c_0\eta$;

(3) $\max_{1\leqslant j\leqslant p} s_j^{-1/2}||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_2 \leqslant c_0\eta$, and $\max_{1\leqslant j\leqslant p} s_j^{-1}||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_1 \leqslant c_0\eta$;

(4) $\max_{1\leqslant j\leqslant p} \widehat{\boldsymbol{\theta}}_j^T\Big(\frac{1}{n}\sum_{i=1}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\widehat{\boldsymbol{\theta}}_j \leqslant 4\xi_p M_2^2/a^4$;

where ξ_1 and ξ_p are the smallest and largest eigenvalue of $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)$, respectively. □

Given the error bounds for $\widehat{\boldsymbol{\Theta}}$, we derive the asymptotic distribution of the debiased estimator $\widetilde{\boldsymbol{\beta}}$, as defined in (3.11), in Theorem 3.2.

Theorem 3.2

Assume the conditions of Lemma 3.2 are satisfied. Let $\Delta_{n,p} = sh^3\sqrt{n\log(p\vee n)} + h^{3/4} + h[\log(p\vee n)]^{3/2} + h\sqrt{\widetilde{s}\log p}$. If $\Delta_{n,p} = o(1)$ and $s\log(p\vee n) \leqslant d_0 n h^5$ for some constant $d_0 > 0$, then

$$\sqrt{n}\big(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big) = \boldsymbol{W} + \boldsymbol{\Delta},$$

for all n sufficiently large, where $\boldsymbol{W} \sim N(\boldsymbol{0}, \boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta})$ with $\boldsymbol{\Lambda} = \mathrm{E}\big\{[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\big\}$

and

$$P(||\boldsymbol{\Delta}||_\infty \geqslant c_0\Delta_{n,p}) \leqslant \exp(-c_1 \log p),$$

for some universal positive constants c_0 and c_1. □

Note that $\Delta_{n,p} = o(1)$ for all sufficiently large n ensures $\eta\widetilde{s} = O(1)$ and $nh^6 \leqslant d_0$ in Lemma 3.2. Theorem 3.2 shows the asymptotic normality of the desparsified estimator $\widetilde{\boldsymbol{\beta}}$. The following corollary establishes uniform validity of the marginal confidence intervals given in (3.17).

Corollary 3.1

Under the conditions of Theorem 3.2,

$$\sup_{||\boldsymbol{\beta}_0||_0 \leqslant s} \max_{1\leqslant j\leqslant p} \sup_{\alpha\in(0,1)} \left| P\Big(\big|\sqrt{n}(\widetilde{\beta}_j - \beta_{0j})\widehat{\Sigma}_{jj}^{-1/2}\big| \leqslant \Phi^{-1}(1-\alpha/2)\Big) - (1-\alpha)\right| = o(1),$$

where $\Phi^{-1}(\cdot)$ is the quantile function of $N(0,1)$. □

Finally, Theorem 3.3 below establishes the validity of the bootstrap procedure introduced in Section 3.2.3 for testing the group hypothesis (3.18). Given a group of variables $\mathcal{G}$, the wild bootstrap test statistic is defined as $\sqrt{n}\max_{j\in\mathcal{G}}|\delta_j^*|$, where $\delta_j^* := \frac{1}{n}\sum_{i=1}^n r_i\{\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j$, $j = 1,\ldots,p$, and $r_1,\cdots,r_n$ are standard normal random variables, independent of the data. Denote $r = \{r_1,\cdots,r_n\}$, and let $w = \{W_1,\cdots,W_n\}$ denote the random sample $W_i = (A_i, \boldsymbol{x}_i, \widetilde{Y}_i)$. Given $0 < \alpha < 1$, recall that the bootstrap critical value for a level-α test is defined as

$$c_{1-\alpha}^* = \inf\big\{t \in \mathbb{R} : P\big(\sqrt{n}\max_{j\in\mathcal{G}}|\delta_j^*| \leqslant t|w\big) \geqslant 1-\alpha\big\} \tag{3.20}$$

Theorem 3.3

Assume the conditions of Theorem 3.2 are satisfied. If $\Delta_{n,p}\sqrt{\log p} = o(1)$ and $d_0\big[\frac{s\log(p\vee n)}{n}\big]^{1/5} \leqslant$

$h \leqslant d_0(\sqrt{\widetilde{s}}\log^2 p)^{-1}$ for some constant $d_0 > 0$, then

$$\sup_{||\boldsymbol{\beta}_0||_0 \leqslant s} \sup_{\alpha \in (0,1)} \left| P\Big(\sqrt{n}\max_{j \in \mathcal{G}} |\widetilde{\beta}_j - \beta_{0j}| \leqslant c^*_{1-\alpha}(\mathcal{G})\Big) - (1-\alpha)\right| = o(1). \qquad \square$$

Theorem 3.3 ensures that the multiplier bootstrap procedure is valid for the simultaneous testing problem (3.18). It is also *honest* in the sense of being valid uniformly over s-sparse models of the form (3.1). It does not require the nonzero components of $\boldsymbol{\beta}_0$ to be well-separated from zero. In particular, the multiple bootstrap procedure does not require the local solution of the profiled estimation to achieve perfect variable selection, which is usually unrealistic in practice.

3.4 Monte Carlo Studies

3.4.1 Algorithm for Estimation

Although it is theoretically possible to compute $\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ via (3.5) for any given $\boldsymbol{\beta}$, the penalized profiled estimating equation can be intensive to solve in high dimension. To increase the computational efficiency, we consider an algorithm based on local linear approximation. Recall that $\boldsymbol{S}_n(\boldsymbol{\beta}) = -n^{-1}\sum_{i=1}^n \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i$. Given a current estimator $\boldsymbol{\beta}^t$ at step t, we update the estimate by

$$\boldsymbol{\beta}^{t+1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p:|\beta_1|=1}{\arg\min} \frac{1}{2n}\sum_{i=1}^n \big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^t)\big]^2 + \lambda||\boldsymbol{\beta}||_1, \quad (3.21)$$

The underlying rationale is that $\boldsymbol{\beta}^{t+1}$ satisfies

$$\boldsymbol{S}_n(\boldsymbol{\beta}^t) + \frac{1}{n}\sum_{i=1}^n \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\big]^2\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t) + \lambda\boldsymbol{\kappa}^{t+1} = \mathbf{0},$$

where $\boldsymbol{\kappa}^{t+1} \in \partial||\boldsymbol{\beta}^{t+1}||_1$. Lemma B.5 in Appendix B.2 implies that with high probability

$$\begin{aligned}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\big]^2\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t)\Big|\Big|_\infty &\leqslant \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\big]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty||\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t||_1 \\ &\leqslant C||\boldsymbol{\beta}^{t+1}-\boldsymbol{\beta}^t||_1,\end{aligned}$$

for some constant $C > 0$. Upon convergence, we will approximately have

$$\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\boldsymbol{\kappa}} = \mathbf{0},$$

where $\widehat{\boldsymbol{\kappa}} \in \partial||\widehat{\boldsymbol{\beta}}||_1$.

An appealing practical property of the algorithm is that the update in step (3.21) can be done very efficiently by treating $\big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) + \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\boldsymbol{x}_i^T\boldsymbol{\beta}^t\big]$ as the working response, and $\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\boldsymbol{x}_i$ as the working covariate vector. This update step can be implemented by many existing algorithms, such as the function "glmnet" in the R package `glmnet` (Friedman et al., 2010). A summary of this algorithm is given in Algorithm 1. In

Algorithm 1 An algorithm for solving the penalized profiled estimating equation.
Input: initial value $\boldsymbol{\beta}^0$, λ, data $\{\boldsymbol{x}_i, \widetilde{Y}_i\}_{i=1}^n$

Set $t = 1$, $\boldsymbol{\beta}^t = \boldsymbol{\beta}^{t-1} = \boldsymbol{\beta}^0$, coef.err = $||\boldsymbol{\beta}^0||_2 + 1$, model.err^{t-1} =model.err^{t-2} = $\mathrm{Var}(\widetilde{Y}_i)$.
while coef.err $> 0.1 * ||\boldsymbol{\beta}^{t-1}||_2$ or model.err^{t-1} > model.err^{t-2} **do**
  $h^t \leftarrow 0.9n^{-1/6}\min\{\mathrm{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t), \mathrm{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)/1.34\}$.
  $w_{ij}^t \leftarrow K\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^t - \boldsymbol{x}_j^T\boldsymbol{\beta}^t}{h^t}\Big)$; $w_{ij}^{\prime t} \leftarrow (h^t)^{-1}K'\Big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}^t - \boldsymbol{x}_j^T\boldsymbol{\beta}^t}{h^t}\Big)$.
  $\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) \leftarrow \frac{\sum_{j\neq i} w_{ij}^t\widetilde{Y}_j}{\sum_{j\neq i} w_{ij}^t}$; $\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) \leftarrow \frac{\sum_{j\neq i} w_{ij}^{\prime t}\widetilde{Y}_j}{\sum_{j\neq i} w_{ij}^t} - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) * \sum_{j\neq i} w_{ij}^{\prime t}$.
  model.err$^t \leftarrow \frac{1}{n}\sum_{i=1}^n[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)]^2$.
  $\boldsymbol{\beta}^{t+1} \leftarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^p:|\beta_1|=1}{\arg\min} \frac{1}{2n}\sum_{i=1}^n[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}^t)\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}^t)]^2 + \lambda||\boldsymbol{\beta}||_1$.
  coef.err$\leftarrow ||\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t||_2$.
  $t \leftarrow t + 1$.
end while
Output $\boldsymbol{\beta}^t$.

implementation, we choose the kernel function $K(\cdot)$ as the distribution function of the standard normal distribution. The bandwidth is set to be $h = 0.9n^{-1/6}\min\{\text{std}(\boldsymbol{x}_i^T\boldsymbol{\beta}), \text{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta})/1.34\}$, as motivated by the suggestion in Silverman (1986), where "std" denotes the standard deviation, and "IQR" denotes the interquartile range. Given a set of candidate tuning parameters $\{\lambda_k\}$ and the corresponding estimators $\widehat{\boldsymbol{\beta}}_{\lambda_k}$, we employ 5-fold cross-validation to select the optimal tuning parameter λ by minimizing $\text{MSE}(\lambda) = n^{-1}\sum_{i=1}^n\{\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda)\}^2$, where $\widetilde{Y}_i = 2(2A_i - 1)Y_i$.

3.4.2 Computation of $\boldsymbol{d}_j(\boldsymbol{\beta}, \eta)$

In Section 3.2.3, we propose a nodewise Dantzig estimator $\boldsymbol{d}_j(\boldsymbol{\beta}, \eta)$, as defined in (3.12), to obtain the approximate inverse of $\nabla \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}})$. This estimator can be solved via a linear programming problem as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\xi}^+,\boldsymbol{\xi}^-\in\mathbb{R}^{p-1}} ||\boldsymbol{\xi}^+||_1 + ||\boldsymbol{\xi}^-||_1 \qquad \text{subject to } \boldsymbol{\xi}^+ \geqslant 0, \boldsymbol{\xi}^- \geqslant 0, \text{ and} \qquad (3.22)\\
\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2 x_{i,k}\boldsymbol{x}_{i,-j}^T(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-) &\geqslant \frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2 x_{i,j}x_{i,k} - \eta, \text{ for all } k \neq j,\\
\frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2 x_{i,k}\boldsymbol{x}_{i,-j}^T(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-) &\leqslant \frac{1}{n}\sum_{i=1}^n[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})]^2 x_{i,j}x_{i,k} + \eta, \text{ for all } k \neq j,
\end{aligned}$$

for any given $j \in \{1, \cdots, p\}$. Then $(\boldsymbol{\xi}^+ - \boldsymbol{\xi}^-)$ is an estimator of $\boldsymbol{d}_j$. In our numerical analysis, we apply the function "lp" in the R package `lpSolve` (Berkelaar and others, 2015) for linear programming.

3.4.3 Monte Carlo Results

We generate random data from the model $Y = \exp(\boldsymbol{x}^T\boldsymbol{\eta}) + (A - \frac{1}{2})\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where the random error $\epsilon \sim N(0, 1)$, the treatment $A \sim$ Bernoulli(0.5), $\boldsymbol{x} = (1, x_1, ..., x_p) = (1, \widetilde{\boldsymbol{x}}^T)^T$, and $\widetilde{\boldsymbol{x}}$ follows a p-dimensional multivariate normal distribution with mean zero

and identity covariance matrix. We set $\boldsymbol{\eta} = (-1, -0.5, 0.5, -0.5, 0, \cdots, 0)^T$ and $\boldsymbol{\beta} = (-2, -0.4, 0.6, -1, -2, 0, \cdots, 0)^T$. As discussed in Section 3.2.1, for the purpose of identifiability, after normalization we have $f_0(u) = 2u$, and $\boldsymbol{\beta}_0 = (-1, -0.2, 0.3, -0.5, -1, 0, \cdots, 0)^T$. We consider sample size $n = 300$ and four different settings with $p = 200, 500, 800, 1000$ in the Monte Carlo experiment.

We first investigate the finite-sample performance of the penalized profiled semiparametric estimator in Section 3.2.2. Table 3.1 reports the average l_1- and l_2-estimation errors, the average number of false negatives (nonzero components incorrectly identified as zero) and false positives (zero components incorrectly identified as nonzero), with their standard deviations in the parentheses. based on 1000 simulation runs. Results in Table 3.1 demonstrate satisfactory performance of the profiled estimator for both the scenarios $p < n$ and $p > n$.

Table 3.1: Performance of the penalized profile least-squares estimator

p	l_1 error	l_2 error	False Negative	False Positive
200	1.06 (0.45)	0.39 (0.12)	0.41 (0.62)	12.13 (11.19)
500	1.21 (0.46)	0.45 (0.14)	0.57 (0.68)	14.73 (14.34)
800	1.29 (0.46)	0.48 (0.14)	0.67 (0.75)	15.73 (16.50)
1000	1.34 (0.50)	0.51 (0.15)	0.71 (0.78)	15.57 (17.40)

Next we investigate the wild bootstrap procedure introduced in Section 3.2.3 for testing the group hypothesis (3.18). Note that for inference, we need to estimate the approximate inverse of $\nabla S_n(\widehat{\boldsymbol{\beta}})$ which involves an additional tuning parameter η. We observe that the inference procedure is not overly sensitive to its choice and fix it at the value $\eta = 15h$ to save computational time. Alternatively, it can also be selected via cross-validation similarly as what has been done for λ selection. For testing hypothesis (3.18), we consider the following six different choices for the groups: $\mathcal{G}_1 = \{6, 7, 8, 9\}$, $\mathcal{G}_2 = \{2, 6, 7, 8, 9\}$, $\mathcal{G}_3 = \{3, 6, 7, 8, 9\}$, $\mathcal{G}_4 = \{2, 3, 6, 7, 8, 9\}$, $\mathcal{G}_5 = \{4, 6, 7, 8, 9\}$ and $\mathcal{G}_6 = \{5, 6, 7, 8, 9\}$. Note that $\mathcal{G}_1$ consists of only zero entries in $\boldsymbol{\beta}_0$, while all the other groups include at least one

non-zero elements. Table 3.2 summarizes the average Type I errors and powers for each scenario, based on 500 Bootstrap samples and 1000 simulation runs.

Table 3.2: Performance of the bootstrap procedure in Section 3.2.3 for simultaneous testing.

p	Type I error	Power				
	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
200	6.5%	55.9%	88.4%	90.5%	100%	100%
500	6.9%	50.1%	87.6%	88.6%	99.8%	100%
800	5.5%	53.2%	86.6%	88.9%	99.7%	100%
1000	7.4%	55.48%	83.1%	86.9%	99.8%	100%

Table 3.2 indicates that type I errors are reasonable controlled for all scenarios. Power performance generally depends on the number and magnitudes of the nonzero components. The hypothesis corresponding to $\mathcal{G}_2$ represents a more challenging situation where the only non-zero element is -0.2, close to 0. The average powers for this case for different values of p are still over 50%. $\mathcal{G}_4$ contains two non-zero elements, one is -0.2 and the other is 0.3, while $\mathcal{G}_3$ has only one non-zero element 0.3. We observe that the magnitude of the smallest nonzero entry has more influence on the power than the number of nonzero coefficients.

3.5 A Real Data Example

We illustrate the application on a clinical data set introduced by Charbonnel et al. (2005). This is a randomized, double-blind, parallel treatment arm, phase III clinical trial to compare the efficacy and safety of pioglitazone versus gliclazide on metabolic control in naive patients with Type 2 diabetes mellitus. This data set we consider contains information on 22 baseline clinical characteristics for 813 individuals with Type 2 diabetes. The patients were randomized into two treatment arms: pioglitazone and gliclazide. Their glycosylated haemoglobin A_{1c} (HbA_{1c}) and fasting plasma glucose (FPG) levels were recorded every four weeks, up to week 52.

The primary efficacy endpoint is the change of HbA_{1c} from baseline to the last available

post-treatment value. We consider all the main effects and two-way interactions in the model. We consider testing the significance of six different groups of variables. Table 3.3 summarizes these six different groups and their respective p-values, based on the bootstrap procedure in Section 3.2.3.

Table 3.3: Real data analysis: evaluation of the significance of different groups of variables

Group	Variables	$p-$value
1	HbA_{1c}, creatinine, fasting insulin, BMI, waist circumference, HomaS	< 0.001
2	all variables in Group 1, and all their two-way interactions	< 0.001
3	HbA_{1c}, fasting insulin, HomaS	< 0.001
4	creatinine, BMI, waist circumference,	0.200
5	HDL-C, LDL-C, total cholesterol, triglycerides	0.180

The first group includes the main effects of six characteristics, which are the baseline average levels for HbA_{1c}, creatinine, fasting insulin, BMI, waist circumference and homeostatic model assessment insulin sensativity (HomaS). The variables in this group are those identified by diabetes experts to be potentially important for optimal treatment regime estimation. The bootstrap procedures suggests a significant p-value (less than 0.001) for this group, which indicates that at one variable in this group is influential for making an optimal personalized decision in the choice of the two treatments. Group 2 augments Group 1 by including all the two-way interaction of these six characteristics, hence includes 21 variables in total). The estimated $p-$value is still less than 0.001. Group 3 and Group 4 are subgroups of Group 1. The third group only includes three main effects: baseline levels of HbA_{1c}, fasting insulin and HomaS, while the fourth group includes the remaining three main effects. The estimated $p-$values suggest that the significant characteristics are among those in Group 3 rather than Group 4. Group 5 consists of four variables: the baseline average levels for the high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol and triglycerides. This group of variables is of interest because Glucose and lipid metabolism are linked to each other in many ways (Parhofer (2015)). Our test suggests that Group 5 does not appear to be influential in

optimal treatment recommendation.

3.6 Discussions

We propose a flexible semi-parametric approach for making honest simultaneous inference about the importance of a group of variables on optimal treatment regime estimation. We develop new statistical theory to overcome the challenges of nonconvexity, high dimensionality and infinite-dimensional nonparametric components. Although the focuses on a randomized trial for ease of presentation, the methods and theory can be extended to observation studies. Assume $\pi(\boldsymbol{x}) = P(A = 1|\boldsymbol{x})$ can be modeled as $\pi(\boldsymbol{x}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a finite-dimensional parameter. Let $\widehat{\xi}$ be an estimate of $\boldsymbol{\xi}$, such as the one based on the popular logistic regression. Under the popular assumption of nounmeasured confounders, a penalized profile semiparametric estimator for $\boldsymbol{\beta}_0$ can be constructed as

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p:|\beta_1|=1}{\arg\min} \frac{1}{2n}\sum_{i=1}^{n}\Big\{\frac{\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta})}{A_i\pi(\boldsymbol{x},\widehat{\boldsymbol{\xi}}) + (1-A_i)(1-\pi(\boldsymbol{x},\widehat{\boldsymbol{\xi}}))}\Big\}^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

The bootstrap inference procedure can be implemented similarly as described in Section 3.2.3.

Chapter 4

A Direct Approach to High-dimensional Error-in-variables Regression

4.1 Introduction

A fundamental assumption in a standard regression model is that the vector of covariates $\boldsymbol{X}$ can be measured exactly and completely. In real applications, sometimes some covariates can only be measured imprecisely or indirectly. For example, in nutrition epidemiology, the measurements of long-term dietary intake is well known to be prone to errors. In genetic studies, measurement error of a phenotypic trait can reduce the power to detect genetic association (Greenwood et al., 2006; Liao et al., 2014). In microarrary studies, the microarray measurements are known to be often subject to various sources of bias (Rocke and Durbin, 2001; Purdom and Holmes, 2005; Tadesse et al., 2005; Boulesteix et al., 2008; Benjamini and Speed, 2012). When such errors-in-variables problem is ignored, biased estimation and misleading inference may result. In the classical asymptotic framework where the number of covariates is small relative to the sample size, this problem was extensively studied, see, for example, the comprehensive introduction of Carroll et al. (2006) and the references therein.

In the classical asymptotic framework where the number of covariates is small and fixed, regularized regression with covariates subject to measurement error has been con-

sidered by Liang and Li (2009); Ma and Li (2010); Zhao and Xue (2010); Sørensen et al. (2015), among others. In this chapter, we are interested in the more challenging setting where the number of covariates can exceed the sample size. Standard high-dimensional regression procedures such as Lasso (Tibshirani, 1996) and Dantzig selector (Candes and Tao, 2007) can be severely biased if measurement error is ignored and may fail to recover the underlying sparsity pattern. Rosenbaum and Tsybakov (2010) observed that the Lasso and Dantzig selector turn out to be extremely unstable in recovering the sparsity pattern even if the measurement error noise level is very small.

Attention was paid to the high-dimensional measurement error problem only recently. Several papers have made important contributions. Loh and Wainwright (2012) systematically studied a general class of problems with corrupted covariates and investigated a non-convex version of Lasso. They proved theoretical results about the statistical error and proposed a projected gradient descent algorithm. Rosenbaum et al. (2013) introduced of a compensated matrix uncertainty (MU) selector that improves their earlier work Rosenbaum and Tsybakov (2010). However, its rate of convergence is suboptimal. The recent paper of Datta and Zou (2017) proposed convex conditioned Lasso which can circumvent the potential nonconvexity in the optimization. Their method first projected the estimate of the covariance matrix of the covariates to the nearest positive semi-definite matrix, and then fitted a standard Lasso regression with the projected matrix. Furthermore, they developed novel results on sign consistency. In another recent paper, Belloni et al. (2017) developed a new estimator that can be written as a second-order cone programming minimization problem. They also made an important theoretical contribution to derive the minimax lower bounds for additive measurement error models and showed that their estimator attains this bound. Other interesting developments in high-dimensional measurement error models include Sørensen et al. (2018), who extended Rosenbaum and Tsybakov (2010) to generalized linear models by devising an efficient algorithm and Rudelson et al. (2017) who developed novel theory and algorithm for a significantly different setting where the

measurement error for each covariate can be a dependent vector across its n observations.

Despite these important developments in methodology and theory, the aforementioned methods all require additional computational efforts, in particular, one or more additional tuning parameters, comparing with those standard approaches in the settings where the complexity of error-in-variables is absent. For example, implementation of the estimator in Loh and Wainwright (2012) requires the knowledge of the L_1 norm of $\boldsymbol{\beta}^*$, the vector of true regression coefficients. Rosenbaum et al. (2013) requires two tuning parameters and the corresponding minimization problem is generally not convex. Although the estimator of Belloni et al. (2017) is guaranteed to be solved numerically in polynomial time, it requires conic programming which is usually slower than linear programming. Datta and Zou (2017) needs an additional step to compute the nearest positive semi-definite matrix projection. Although an efficient ADMM algorithm was proposed for this purpose, this step still demands extra tuning parameters and computational time.

Building on this recent work, in particular, extending the novel work of Loh and Wainwright (2012), we propose an alternative estimator that enjoys both computational convenience and desirable statistical properties. It has the same computational efficiency of standard Dantzig estimator in the non-contamination case and requires no additional tuning parameter besides the one for the standard Dantzig estimator. It can be implemented using any existing software for linear programming. Our numerical studies demonstrate that it substantially reduces the computational time compared with existing methods. Theoretically, we derive the estimation error bound. For additive measurement error models, it achieves the minimax efficiency bound proved in Belloni et al. (2017).

The rest of this chapter is organized as follows. Section 4.2 introduces the proposed new estimator. Section 4.3 derives the nonasymptotic error bound and illustrates the theory in three representative examples. Section 4.4 demonstrates the finite sample performance of the new method in Monte Carlo studies. Section 4.5 concludes the chapter.

4.2 Methodology

4.2.1 Problem setup

We consider the following linear regression models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \tag{4.1}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the vector of responses, $\boldsymbol{X} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T \in \mathbb{R}^{n\times p}$ is the matrix of covariates, $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T$ is the noise vector independent of $\boldsymbol{X}$. In the high dimensional setting, the number of covariates p can exceed the sample size n. It is assumed that $\boldsymbol{\beta}^*$ is sparse with sparsity size $s = ||\boldsymbol{\beta}^*||_0$, where s is allowed to increase with n but is assumed to be small compared with n.

In this chapter, we adopt the general setup as in Loh and Wainwright (2012). Instead of observing a clean covariate matrix $\boldsymbol{X} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T \in \mathbb{R}^{n\times p}$, we observe a matrix $\boldsymbol{Z} = (\boldsymbol{z}_1^T, ..., \boldsymbol{z}_n^T)^T \in \mathbb{R}^{n\times p}$, which can be considered as a corrupted version of $\boldsymbol{X}$, due to measurement error or missing data. Conditional on $\boldsymbol{x}_i$, $\boldsymbol{z}_i$ is assumed to have the conditional distribution

$$\boldsymbol{z}_i \sim \mathbb{Q}(\cdot|\boldsymbol{x}_i) \text{ for } i = 1, ..., n. \tag{4.2}$$

Our goal is to reliably estimate the vector of unknown regression coefficients $\boldsymbol{\beta}^*$ based on $\boldsymbol{Z}$ and $\boldsymbol{y}$.

An important special case is the additive measurement error model. In this case, instead of $\boldsymbol{X}$, we observe $\boldsymbol{Z}$ such that

$$\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}, \tag{4.3}$$

with $\boldsymbol{W}$ being the additive measurement error independent of $\boldsymbol{X}$.

4.2.2 Proposed method

Our proposed estimator adapts the Dantzig estimator (Candes and Tao, 2007) to the setting of covariates with measurement error. To motivate the new estimator, we first consider the ideal setting, where we directly observe $\boldsymbol{X}$. The Dantzig procedure estimates $\boldsymbol{\beta}^*$ by minimizing $||\boldsymbol{\beta}||_1$ subject to the constraint

$$\left|\left|n^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right|\right|_\infty < r, \tag{4.4}$$

where $r > 0$ is a tuning parameter, and $||\cdot||_\infty$ denotes the L_∞ norm. Bickel et al. (2009) showed that it enjoys similar statistical properties as the Lasso.

Let $\boldsymbol{\gamma} = n^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\boldsymbol{\Gamma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. In the measurement error case, we do not directly observe $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$. We replace $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ by their unbiased estimators $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$, respectively, and consider the following feasible estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min ||\boldsymbol{\beta}||_1 \quad \text{such that } ||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}||_\infty < r. \tag{4.5}$$

In the next section, we will show the near oracle property of this corrected estimator under some mild assumptions. For the additive measurement model (4.3), we could use

$$\widehat{\boldsymbol{\Gamma}}_{\text{add}} = n^{-1}\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{\Sigma}_w, \qquad \text{and} \qquad \widehat{\boldsymbol{\gamma}}_{\text{add}} = n^{-1}\boldsymbol{Z}^T\boldsymbol{y}, \tag{4.6}$$

where following Loh and Wainwright (2012) we assume that the rows of $\boldsymbol{W}$ are i.i.d random vectors with mean zero and known covariance matrix $\boldsymbol{\Sigma}_w$.

The optimization problem in (4.5) can be recast as a linear programming problem, as follows. Write $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, where $\boldsymbol{\beta}^+ = (\beta_1^+, \ldots, \beta_p^+)^T$ and $\boldsymbol{\beta}^- = (\beta_1^-, \ldots, \beta_p^-)^T$ are the vectors of the positive and negative parts of the entries of $\boldsymbol{\beta}$, respectively. Then solving

for $\widehat{\boldsymbol{\beta}}$ is equivalent to

$$\begin{aligned} & \min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-} \Big\{ \sum_{i=1}^{n} (\beta_i^+ + \beta_i^-) \Big\} \\ \text{subject to} \quad & \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^+ - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^- \leqslant r\mathbf{1} + \widehat{\boldsymbol{\gamma}} \\ & -\widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^+ + \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^- \leqslant r\mathbf{1} - \widehat{\boldsymbol{\gamma}} \\ & \beta_i^+ \geqslant 0, \beta_i^- \geqslant 0;\ i = 1, 2, \cdots, p, \end{aligned} \tag{4.7}$$

where $\mathbf{1}$ denotes a n-vector of ones. This can be easily implemented using any existing software for linear programming. For instance, the recently developed R package fastclime can be applied to solve large-scale linear programs and efficiently calculate the full piecewise-linear regularization path.

Remark 4.1

Although our proposed estimator was motivated and related to some recent work, there exist substantial difference. The above estimator was partially motivated by Loh and Wainwright (2012), who proposed a modified Lasso estimator

$$\min_{||\boldsymbol{\beta}||_1 \leqslant b_0\sqrt{s}} \big\{ \frac{1}{2}\boldsymbol{\beta}^T\widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta} - \langle \widehat{\boldsymbol{\gamma}}, \boldsymbol{\beta} \rangle + \lambda||\boldsymbol{\beta}||_1 \big\} \tag{4.8}$$

where b_0 is a suitable constant, and s is the sparsity size of $\boldsymbol{\beta}^*$. However, their estimator requires the knowledge of $||\boldsymbol{\beta}^*||_1$. The objective function can be nonconvex and requires an alternative algorithm rather than standard Lasso. In contrast, for the proposed new estimator, the objective function is convex. Furthermore, the constraints in a linear program problem always produce a convex polytope as the feasible region. The new estimator can be solved directly by any standard linear programming software. Upon first sight, out new estimator appear to be related to the compenstated compensated matrix uncertainty selector of Rosenbaum et al. (2013) by setting the μ in their equation (3) to 0. However,

this connection to our proposed new estimator is actually superficial. Careful reading of their paper reveals that $\mu = 0$ only applies to the deterministic (noiseless) case (i.e., no noise in the underlying regression model or the contamination model). The most interesting case in practice is the stochastic error case, where there is a random noise error in both the underlying regression model and the contamination model. For the stochastic error case, their main theory requires $\mu > 0$ and does not apply to our new estimator. Furthermore, as pointed out in Belloni et al. (2017), its L_2 error bound is suboptimal. Finally, the proposed estimator is related to Belloni et al. (2017), who constructed a conic-programming-based estimator for the additional noise case. Their estimator is obtained by solving

$$\begin{gathered} \text{minimize } ||\boldsymbol{\beta}||_1 + \lambda t \text{ over } (\boldsymbol{\beta}, t) \text{ such that} \\ \boldsymbol{\beta} \in \mathbb{B}, t \in \mathbb{R}^+, \ ||n^{-1}\boldsymbol{Z}^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \widehat{\boldsymbol{D}}\boldsymbol{\beta}||_\infty \leqslant \mu t + \tau, \ ||\boldsymbol{\beta}||_2 \leqslant t \end{gathered} \tag{4.9}$$

where λ, τ and μ are all positive tuning parameters, and $\mathbb{B} \subseteq \mathbb{R}^p$ defines the range for $\boldsymbol{\beta}$. In general, linear programming is faster than conic programming; and the proposed new estimator has only one tuning parameter. □

Remark 4.2

The new estimator in (4.5) depends on a tuning parameter r. We can choose r by a k-fold cross-validation:

$$\tilde{r} = \arg\min_r \sum_{k=1}^{K} \{\widehat{\boldsymbol{\beta}}_k(r)^T \widehat{\boldsymbol{\Gamma}}_k \widehat{\boldsymbol{\beta}}_k(r) - 2\widehat{\boldsymbol{\gamma}}_k^T \widehat{\boldsymbol{\beta}}_k(r)\}, \tag{4.10}$$

where $\widehat{\boldsymbol{\beta}}_k(r)$, $\widehat{\boldsymbol{\Gamma}}_k$ and $\widehat{\boldsymbol{\gamma}}_k$ are computed when the kth part of the data are left out. Alternatively, we can use the cross-validation proposed in Datta and Zou (2017) which chooses

$$\hat{r} = \arg\min_r \sum_{k=1}^{K} \{\widehat{\boldsymbol{\beta}}_k(r)^T (\widehat{\boldsymbol{\Gamma}}_k)_+ \widehat{\boldsymbol{\beta}}_k(r) - 2\widehat{\boldsymbol{\gamma}}_k^T \widehat{\boldsymbol{\beta}}_k(r)\}. \tag{4.11}$$

Note that (4.11) replaces the $\widehat{\mathbf{\Gamma}}_k$ in (4.10) by $(\widehat{\mathbf{\Gamma}}_k)_+$, the projected positive definite matrix, due to the concern that the negative eigenvalues of $\widehat{\mathbf{\Gamma}}_k$ can potentially lead to an unbounded objective function. However, we observe that $\widehat{\mathbf{\Gamma}}_k$ satisfies the restricted eigenvalue condition (condition (4.13) in Section 4.3.1) with high probability. This together with Lemma 4.1 below implies that the unboundedness usually does not happen in practice. We examine both cross-validation methods in simulations and find they have similar performance while (4.10) is more computationally efficient. □

4.3 Statistical theory

4.3.1 L_1 and L_2 estimation error bounds

In this subsection, we derive the L_1 and L_2 error bounds of the proposed estimator $\widehat{\boldsymbol{\beta}}$ in (4.5). We first define the following two events.

Let event Ω_{n1} be the event that there exists a function $\varphi(\mathbb{Q}, \sigma_\epsilon)$ that depends on the source of the contamination noise $\mathbb{Q}$ in (4.2) and the standard deviation σ_ϵ of $\boldsymbol{\epsilon}$ in (4.1) such that

$$||\widehat{\boldsymbol{\gamma}} - \widehat{\mathbf{\Gamma}}\boldsymbol{\beta}^*||_\infty \leqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}. \tag{4.12}$$

Let event Ω_{n2} be the event that there exist $\alpha_1 > 0$ and $\tau(n, p) > 0$ such that

$$\boldsymbol{\theta}^T\widehat{\mathbf{\Gamma}}\boldsymbol{\theta} \geqslant \alpha_1||\boldsymbol{\theta}||_2^2 - \tau(n, p)||\boldsymbol{\theta}||_1^2 \tag{4.13}$$

for all $\boldsymbol{\theta} \in \mathbb{R}^p$.

Remark 4.3

These two events are the same as those in Loh and Wainwright (2012). In particular, event

Ω_{n1} holds if the following two conditions hold simultaneously:

$$||\widehat{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_x\boldsymbol{\beta}^*||_\infty \leqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}} \quad \text{and} \quad ||(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Sigma}_x)\boldsymbol{\beta}^*||_\infty \leqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}}, \tag{4.14}$$

where $\boldsymbol{\Sigma}_x$ is the covariance matrix of each clean covariate vector $\boldsymbol{x}_i$. Condition (4.13) was referred to as a lower restricted eigenvalue (Lower-RE) condition for the estimated covariance matrix $\widehat{\boldsymbol{\Gamma}}$. It was verified in Loh and Wainwright (2012) that these conditions are satisfied with high probability for several important applications. It can be seen from Lemma 4.1 below that condition (4.13) can be relaxed to require the lower bound to hold on a restricted set $\{\boldsymbol{\nu} \in R^p : ||\boldsymbol{\nu}_{T^C}||_1 \leqslant ||\boldsymbol{\nu}_T||_1\}$, where $T = \{j : \beta_j^* \neq 0, j = 1, ..., p\}$ denotes the index set of nonzero coefficients. For an arbitrary vector $\boldsymbol{\nu}$, $\boldsymbol{\nu}_T$ denote the subvector of $\boldsymbol{\nu}$ with entries corresponding to indice in T; and let $\boldsymbol{\nu}_{T^c}$ be defined similarly, where T^c denotes the complement of set T. □

A nice property of the proposed new estimator $\widehat{\boldsymbol{\beta}}$ is that it immediately satisfies a cone constraint under some very mild conditions.

Let $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Let $\boldsymbol{\delta}_T$ denote the subvector of $\boldsymbol{\delta}$ with entries corresponding to indice in T, and $\boldsymbol{\delta}_{T^c}$ be defined similarly, where T^c denotes the complement of set T.

Lemma 4.1

On Ω_{n1}, if $r \geqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\log p/n}$, then we have

$$||\boldsymbol{\delta}_{T^C}||_1 \leqslant ||\boldsymbol{\delta}_T||_1$$

with the deviation bound (4.12) satisfied. □

Proof of Lemma 4.1 First, we show that $\boldsymbol{\beta}^*$ satisfies the constraint $||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^*||_\infty < r$. By the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$||\boldsymbol{\beta}^*||_1 \geqslant ||\widehat{\boldsymbol{\beta}}||_1 = ||\boldsymbol{\beta}^* + \boldsymbol{\delta}||_1 = ||\boldsymbol{\beta}^* + \boldsymbol{\delta}_T||_1 + ||\boldsymbol{\delta}_{T^C}||_1 \geqslant ||\boldsymbol{\beta}^*||_1 - ||\boldsymbol{\delta}_T||_1 + ||\boldsymbol{\delta}_{T^C}||_1.$$

This implies $||\boldsymbol{\delta}_{T^C}||_1 \leqslant ||\boldsymbol{\delta}_T||_1$. □

Theorem 4.1

On $\Omega_{n1} \cap \Omega_{n2}$, if $8s\tau(n,p) \leqslant \alpha_1$ and $r \geqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n}$, then the estimator $\widehat{\boldsymbol{\beta}}$ in (4.5) satisfies

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leqslant \frac{4\sqrt{s}}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}} + r) \tag{4.15}$$

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leqslant \frac{8s}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\frac{\log p}{n}} + r) \tag{4.16}$$

□

The error bounds in Theorem 4.1 are non-asymptotic. The proof of Theorem 4.1 is given in Appendix C.

In Examples 1-3 below, we will build on the results of Loh and Wainwright (2012) and provide more specific error bounds in some important applications. In particular, these bounds are shown to hold with probability going to one exponentially fast. In these examples, we assume $\boldsymbol{X}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}_x, \sigma_x^2)$, that is the rows $\boldsymbol{x}_i$'s are sampled independently from a p-dimensional zero-mean distribution with covariance $\boldsymbol{\Sigma}_x$, and for any unit vector $\boldsymbol{u} \in \mathbb{R}^p$, the random variable $\boldsymbol{u}^T\boldsymbol{x}_i$ is sub-Gaussian with parameter at most σ_x. We also assume that $\boldsymbol{\epsilon}$ is a sub-Gaussian random random vector with parameter σ_ϵ^2.

4.3.2 Example 1: Additive measurement error model

Recall that in the additive measurement error model, instead of $\boldsymbol{X}$, we observe $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, where $\boldsymbol{W}$ is the additive measurement errors independent of $\boldsymbol{X}$. We assume $\boldsymbol{W}$ is sub-Gaussian with parameters $(\boldsymbol{\Sigma}_w, \sigma_w^2)$. Consider $\widehat{\boldsymbol{\Gamma}}_{\text{add}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{\Sigma}_w$ and $\widehat{\boldsymbol{\gamma}}_{\text{add}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{y}$ as in (4.6). Theorem 4.1 can be applied with $\alpha_1 = \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_x)$,

$$\tau(n,p) = d_0\lambda_{\min}(\boldsymbol{\Sigma}_x)\max\Big(\frac{\sigma_z^4}{\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}, 1\Big)\frac{\log p}{n},$$

and $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, where $\lambda_{\min}(\boldsymbol{\Sigma}_x)$ denotes the smallest eigenvalue of $\boldsymbol{\Sigma}_x$, $c_0 = d_1\sigma_z(\sigma_w + \sigma_\epsilon)$, $\sigma_z^2 = \sigma_x^2 + \sigma_w^2$, and d_0, d_1 are positive constants. This leads to the following more explicit error bound for the additive measurement error model.

Corollary 4.1

If $n \geqslant c' \max\{\frac{\sigma_x^4}{\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}, 1\}s\log p$ with some positive constant c', for $r \geqslant c_0||\boldsymbol{\beta}^*||_2\sqrt{(\log p)/n}$, we have

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leqslant \frac{4\sqrt{s}}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1\exp(-c_2\log p), \quad (4.17)$$

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leqslant \frac{8s}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1\exp(-c_2\log p), \quad (4.18)$$

for some positive constants c_0, c_1 and c_2. □

Remark 4.4

Consequently, we have $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leqslant c^*\frac{\sqrt{s}}{\alpha_1}||\boldsymbol{\beta}^*||_2\sqrt{(\log p)/n}$ and $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leqslant c^*\frac{s}{\alpha_1}||\boldsymbol{\beta}^*||_2\sqrt{(\log p)/n}$ with high probability, for some positive constant c^*. This matches the minimax rate derived in Belloni et al. (2017) for the Gaussian setting. □

4.3.3 Example 2: Missing data model

In model (4.1), consider the case that the entries of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are missing at random. Let $\boldsymbol{\rho} = (\rho_1, ..., \rho_p)^T$ be the missing probability vector. The observed covariates can be denoted as

$$z_{ij} = \begin{cases} x_{ij}, & \text{with probability } 1-\rho_j \\ 0, & \text{otherwise} \end{cases}$$

where $z_{ij} = 0$ means the jth covariate of the ith observation is missing. We could construct the unbiased estimators as in Loh and Wainwright (2012):

$$\widehat{\boldsymbol{\Gamma}}_{\text{miss}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} \oslash \boldsymbol{M}, \qquad \text{and} \qquad \widehat{\boldsymbol{\gamma}}_{\text{miss}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{y} \oslash (\boldsymbol{1}-\boldsymbol{\rho}), \tag{4.19}$$

where $\oslash$ denotes element-wise division, and $\boldsymbol{M}$ satisfies:

$$M_{ij} = \begin{cases} (1-\rho_i)(1-\rho_j), & \text{if } i \neq j \\ 1-\rho_i, & \text{if } i = j \end{cases}.$$

Remark 4.5

In the special case that all covariates are missing with the same probability ρ, then we can simplify the unbiased estimators as:

$$\widehat{\boldsymbol{\Gamma}}_{\text{miss}} = \frac{1}{n(1-\rho)^2}\{\boldsymbol{Z}^T\boldsymbol{Z} - \rho\,\text{diag}(\boldsymbol{Z}^T\boldsymbol{Z})\}, \qquad \text{and} \qquad \widehat{\boldsymbol{\gamma}}_{\text{miss}} = \frac{1}{n(1-\rho)}\boldsymbol{Z}^T\boldsymbol{y} \qquad \square$$

Theorem 4.1 can be applied with $\alpha_1 = \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_x)$, $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, and $\tau(n,p) = d_0\lambda_{\min}(\boldsymbol{\Sigma}_x)\max(\frac{1}{(1-\rho_{\max})^4} * \frac{\sigma_x^4}{\lambda^2_{\min}(\boldsymbol{\Sigma}_x)}, 1)\frac{\log p}{n}$, where $c_0 = d_1\frac{\sigma_x}{1-\rho_{\max}}(\sigma_\epsilon + \frac{\sigma_x}{1-\rho_{\max}})$ and d_0, d_1 are positive constants.

Corollary 4.2

If $n \geqslant c' \max\{\frac{1}{(1-\rho_{\max})^4} * \frac{\sigma_x^4}{\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}, 1\} s \log p$, with some positive constant c', for $r \geqslant c_0 ||\boldsymbol{\beta}^*||_2 \sqrt{(\log p)/n}$, we have

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leqslant \frac{4\sqrt{s}}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1 \exp(-c_2 \log p), \quad (4.20)$$

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leqslant \frac{8s}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1 \exp(-c_2 \log p), \quad (4.21)$$

for some positive constants c_0, c_1 and c_2. □

4.3.4 Example 3: Multiplicative noise model

In this case, we can write the observed covariate matrix $\boldsymbol{Z}$ as $\boldsymbol{Z} = \boldsymbol{X} \odot \boldsymbol{U}$, where $\odot$ denotes element-wise multiplication, and $\boldsymbol{U}$ is assumed to be independent of $\boldsymbol{X}$. This is a generalization of Example 2. In the missing data case, we can regard the missing data as being resulted from corruption with independent multiplicative binary noises u_{ij}. In other words, the noise matrix $\boldsymbol{U}$ has elements $u_{ij} = 0$ if $x_{ij} = 0$, and $u_{ij} = 1$, otherwise. For different observations, the noise vectors $\boldsymbol{u}_1, ..., \boldsymbol{u}_n$, where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^T$, $i = 1, \ldots, n$, are assumed to be independent. We may consider other forms for the noise matrix $\boldsymbol{U}$. In the following, we assume that the multiplicative noise $\boldsymbol{U}$ is bounded by K, i.e., $|u_{ij}| < K$ always holds for any $i = 1, ..., n,\ j = 1, ..., p$.

Define

$$\boldsymbol{M} = E(\boldsymbol{u}_1 \boldsymbol{u}_1^T), \qquad \text{and} \qquad \boldsymbol{l} = E(\boldsymbol{u}_1).$$

To avoid singularity, we assume that for the covariance matrix $\boldsymbol{M} = (m_{ij})_{p\times p}$ and the expectation vector $\boldsymbol{l} = (l_1, ..., l_p)^T$, all elements are positive, i.e., $m_{ij} > 0$ for all $i, j = 1, ..., p$, and $l_i > 0$ for all $i = 1, ..., p$. Then we could construct the unbiased estimators as

in Loh and Wainwright (2012):

$$\widehat{\mathbf{\Gamma}}_{\text{multi}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} \oslash \boldsymbol{M}, \qquad \text{and} \qquad \widehat{\boldsymbol{\gamma}}_{\text{multi}} = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{y} \oslash \boldsymbol{l}, \tag{4.22}$$

Theorem 4.1 can be applied with $\alpha_1 = \frac{1}{2}\lambda_{\min}(\mathbf{\Sigma}_x)$, $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, and $\tau(n,p) = d_0\lambda_{\min}(\mathbf{\Sigma}_x)\max(\frac{K^4\sigma_x^4}{m_{\min}^2\lambda_{\min}^2(\mathbf{\Sigma}_x)}, 1)\frac{\log p}{n}$, where $c_0 = d_1\frac{K\sigma_x}{l_{\min}}(\frac{m_{\min}+Kl_{\min}}{m_{\min}}\sigma_x + \sigma_\epsilon)$, $m_{\min}$ is the minimum element of the covariance matrix $\boldsymbol{M}$, and $l_{\min}$ is the minimum element of the expectation vector $\boldsymbol{l}$, and d_0, d_1 are positive constants.

Corollary 4.3

If $n \geqslant c' \max\{\frac{K^4\sigma_x^4}{m_{\min}^2\lambda_{\min}^2(\mathbf{\Sigma}_x)}, 1\}s\log p$, with some positive constant c', for $r \geqslant c_0||\boldsymbol{\beta}^*||_2\sqrt{(\log p)/n}$, we have

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2 \leqslant \frac{4\sqrt{s}}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1\exp(-c_2\log p), \tag{4.23}$$

$$P\big(||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leqslant \frac{8s}{\alpha_1}(\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)\big) \geqslant 1 - c_1\exp(-c_2\log p), \tag{4.24}$$

for some positive constants c_0, c_1 and c_2. □

Remark 4.6

In the missing data case, consider $K = 1$, $l_{\min} = 1 - \rho_{\max}$ and $m_{\min} = (1 - \rho_{\max})^2$. Then we reduce Corollary 4.3 to Corollary 4.2.

4.4 Monte Carlo studies

In the Monte Carlo studies, we compare the proposed estimator with four alternative methods: the convex conditioned Lasso in Datta and Zou (2017)(denoted by CoCo), the conic-programming based estimator in (Belloni et al., 2017) (denoted by Conic), the Oracle Dantzig method (standard Dantzig method applied to the clean covariate matrix $\boldsymbol{X}$) and

the Naive Dantzig method (standard Dantzig method directly applied to $\boldsymbol{Z}$). It is worth mentioning that (Datta and Zou, 2017) has compared with the estimator in Loh and Wainwright (2012) and showed they have competitive performance; (Belloni et al., 2017) has compared with the estimator in (Rosenbaum and Tsybakov, 2013) and demonstrated Conic has slightly better performance. Hence, we do not directly include Loh and Wainwright (2012) and (Rosenbaum and Tsybakov, 2013) in our comparision. We implemented Conic using the Matlab code provided on the authors' website and implemented CoCo using the R code the authors shared with us.

Example 1 (Regression with additive measurement error). Our simulation setup is taken from Belloni et al. (2017). We generate random data from the model $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ from the model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^* + \epsilon_i$, $i = 1, \ldots, n$, where $\boldsymbol{\beta}^* = (1, 1, 1, 1, 1, 0, ..., 0)^T \in \mathbb{R}^p$, ϵ_i are independent $\mathcal{N}(0, \sigma^2)$ random variables. The clean covariates x_i are independently generated from a p-dimensional multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_x$ with $(\boldsymbol{\Sigma}_x)_{ij} = \rho^{|i-j|}$, $1 \leqslant i, j \leqslant p$. We then generate the observed covariates $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, where $\boldsymbol{W} = (\boldsymbol{w}_i)_{i=1,\ldots,n}$ are the measurement errors independent of $\boldsymbol{X}$. We generate $\boldsymbol{w}_i's$ independently from a p-dimensional multivariate normal distribution with zero mean and covariance matrix $\tau^2 I_{p\times p}$. As in Belloni et al. (2017), we set $\sigma = 0.128$, $\rho = 0.25$ and $\tau = 0.45$. We consider two different settings for sample size n and dimension p: $n = 300$, $p = 100$; and $n = 200$, $p = 300$.

Both Oracle Dantzig and Naive Dantzig used 5-fold cross-validation to select the constraint parameter which minimizes the prediction errors. Since our setup is the same as that in Belloni et al. (2017), we use the tuning parameters reported in their paper for Conic. In this way, we can be as fair as possible for all of these methods. For CoCo, the tuning parameter is chosen using the 5-fold corrected cross-validation proposed in their paper. For New, we choose the tuning parameter using the 5-fold cross-validation using (4.10) (denoted by New (CV_1) and (4.11) (denoted by New (CV_2).

The simulation results based on 100 runs are summarized in Table 4.1. In the table, *C*

Table 4.1: Simulation results for Example 1

Setting	Method	C	IC	PE(se)	MSE(se)
1	New (CV_1)	5	8.33	0.066 (0.029)	0.060 (0.027)
	New(CV_2)	5	8.73	0.068 (0.027)	0.061 (0.028)
	CoCo	5	15.74	0.070 (0.030)	0.068 (0.033)
	Conic(0.5)	5	7.80	0.063 (0.024)	0.067 (0.027)
	Conic(0.75)	5	9.97	0.070 (0.026)	0.075 (0.029)
	Conic(1)	5	10.14	0.070 (0.026)	0.075 (0.029)
	Oracle Dantzig	5	0.02	0.002 (0.001)	0.001 (0.001)
	Naive Dantzig	5	16.08	0.244 (0.049)	0.181 (0.036)
2	New (CV_1)	5	15.32	0.127 (0.050)	0.112 (0.043)
	New (CV_2)	5	11.50	0.130 (0.052)	0.112 (0.046)
	CoCo	5	18.10	0.120 (0.046)	0.109 (0.041)
	Conic(0.5)	5	7.11	0.098 (0.037)	0.098 (0.039)
	Conic(0.75)	5	11.83	0.107 (0.036)	0.114 (0.043)
	Conic(1)	5	16.30	0.126 (0.036)	0.133 (0.046)
	Oracle Dantzig	5	0.26	0.004 (0.002)	0.003 (0.001)
	Naive Dantzig	5	26.53	0.362 (0.087)	0.271 (0.063)

denotes the average number of nonzero coefficients correctly identified as nonzero; *IC* denotes the average number of zero coefficients incorrectly identified as nonzero; *PE* denotes the prediction error; and *MSE* is the mean squared error. More specifically,

$$\text{PE}(\widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{\Sigma}_x (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \quad \text{and} \quad \text{MSE}(\widehat{\boldsymbol{\beta}}) = ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2.$$

And the standard errors for them in the table are calculated by Bootstrap.

The Oracle Dantzig can be considered as the gold standard in the ideal situation where clean covariates are directly observed; while the Naive Dantzig completely ignores data corruption. As expected, we observe in Table 4.1 that Naive Dantzig is severely biased. The proposed new estimator (with tuning parameter chosen either by CV_1 or CV_2) has overall competitive performance comparing with both CoCo and Conic in terms of both PE and MSE. It is able to correctly identify all nonzero coefficients while maintaining a relatively small number of false positives. In terms of computational speed, the improvement in

Table 4.2: Simulation results for Example 2

Setting	Method	C	IC	PE(se)	MSE(se)
1	New (CV_1)	5	14.14	0.115 (0.062)	0.115 (0.060)
	New (CV_1)	5	13.88	0.116 (0.063)	0.115 (0.060)
	CoCo	5	14.70	0.116 (0.062)	0.117 (0.063)
	Oracle Dantzig	5	0.00	0.002 (0.001)	0.001 (0.001)
	Naive Dantzig	5	15.71	0.734 (0.143)	0.518 (0.102)
2	New (CV_1)	5	19.74	0.193 (0.072)	0.182 (0.072)
	New (CV_2)	5	13.73	0.185 (0.072)	0.170 (0.070)
	CoCo	5	15.41	0.172 (0.068)	0.160 (0.066)
	Oracle Dantzig	5	0.31	0.004 (0.002)	0.003 (0.001)
	Naive Dantzig	5	27.71	0.871 (0.155)	0.622 (0.106)

the high-dimensional setting is substantial. For fair comparison, we fixed all tuning parameters: for $p = 300$, the proposed estimator costs about 0.23 seconds for one simulation run; while Conic needs about 4 seconds and CoCo needs around 9 seconds. The proposed estimator only has one tuning parameter, suggesting further savings in computing time when cross-validation is used to choose the tuning parameter.

Example 2 (Regression with multiplicative measurement error). In this example, the covariate matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{y}$ are generated the same way as in Example 1. However, the observed covariate matrix in this setup is $\boldsymbol{Z} = \boldsymbol{X} \odot \boldsymbol{U}$, where $\boldsymbol{U} = (u_{ij})_{n\times p}$ is the measurement error matrix. Each element u_{ij} follows log-normal distribution: $\log(u_{ij}) \sim \mathcal{N}(0, \tau^2)$ with $\tau = 0.5$. Since Belloni et al. (2017) was focused on additive measurement error model, it is not included in this example. We compare New with CoCo, Oracle Danztig and Naive Dantzig. As Example 1, 5-fold corrected cross-validation was applied to select the tuning parameter for New and CoCo. The simulation results are summarized in Table 4.2.

Similarly as in Example 1, naive Dantzig is severely biased; the proposed new estimator (with tuning parameter chosen either by CV_1 or CV_2) has overall competitive

performance. For fixing all tuning parameters and $p = 300$, the proposed estimator costs about 0.2 seconds for one simulation run; while CoCo needs around 9 seconds.

4.5 Conclusion and discussions

Our work introduces a new approach for estimating a high-dimensional error-in-variable regression model. The proposed new estimator enjoys the computational advantage of classical Dantzig estimator with only one tuning parameter and can be solved by standard linear programming software. It has an L_2 estimation error bound with the same near oracle rate of Lasso or Dantzig with covariates being observed without error. For additive measurement error, this bound achieves the minimax rate.

Similarly as Loh and Wainwright (2012), Datta and Zou (2017) and Belloni et al. (2017), our method has not addressed the problem of estimating the variance of the measurement errors. This is itself a challenging problem in the high-dimensional regimes. However, in some important cases, a data-driven estimator of the measurement error variance is feasible. For example, Rosenbaum et al. (2013) described such a scenario of missing data problem. In general, estimating the the variance of the measurement errors requires repeated measurements. This will be an interesting topic for future research.

Chapter 5

Conclusion and Discussion

In this dissertation we have studied statistical inference problems about the optimal treatment regime in two different setting, and the high-dimensional error-in-variables regression estimation problem, which is relevant to the precision medicine analysis.

Firstly, we have proposed a smoothed robust estimator that directly targets estimating the parameters corresponding to the Bayes decision rule for estimating the optimal treatment regime. This estimator is shown to have an asymptotic normal distribution. Furthermore, it is proved that a resampling procedure provides asymptotically accurate inference for both the parameters indexing the optimal treatment regime and the optimal value function.

For high-dimensional inference, we have developed new tools to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigated simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference for the effect of a group of variables on optimal decision making. The advantage of honest inference is that it does not require the initial estimator to achieve perfect model selection and does not require the

zero and nonzero effects to be well-separated.

We also developed new algorithms with solid theoretical foundation for both of the above two methods. Our numerical results suggest satisfactory performance of the new methods. Meanwhile, though both methods focus on a randomized trial for ease of presentation, they can be extended to observation studies, see discussions in Section 2.6.1 and Section 3.6.

In addition, we proposed a new direct, computationally simple approach for estimating a high-dimensional error-in-variable regression model, which enjoys the same computational convenience of standard Dantzig estimator with perfectly measured covariates and requires no additional tuning parameter. The new estimator can be easily implemented using any existing software for linear programming. Theoretically, we prove that it achieves an estimation error bound with a near-optimal minimax rate.

As discussed in Section 4.5, this new approach requires the knowledge of variance structure of the measurement errors. Estimating the variance structure is itself a challenging problem in the high-dimensional regimes. An interesting topic for future research is that estimating the variance of the measurement errors with repeated measurements.

There are still several attractive questions we can go further in precision medicine. One of them is the estimation and inference for dynamic treatment regimes (DTRs). This is another main challenge in precision medicine. Compared with as searching the optimal treatment regime at any point in time conditions, the sequential decision rules, i.e., dynamic treatment regimes, for individual patients that can adapt over time to an evolving illness make more sense in clinical medicine. DTRs take both diversity across patients and diversity over time within each patient into consideration. A number of methods have been proposed to estimate the optimal DTRs. However, model assumptions still play an important role in these methods. In my future research, developing a robust approach to search the optimal DTR, and figuring out the significant features among a large scale of characteristics is a critical topic.

Another interesting point is to extend our new approach for high-dimensional error-in-variables regressions estimation to observational studies in precision medicine. It is straightforward to apply our proposed method to randomized clinical trials with measurement errors in precision medicine. However, when taking into account the propensity score function estimation in observational studies, much more obstacles are involved. First of all, we need to develop a new method for high-dimensional generalized linear regressions with measurement errors. Nonlinear link function leads to difficulties in eliminating the influence of measurement errors. Next, due to the possibility of the model misspecification for the propensity score function, it is also worthwhile to consider a doubly robust estimator in high-dimensional precision medicine analysis. Finally, if it is feasible, we are also interested in making inference about the optimal treatment regime in observational studies. My next step of the research will focus on these impressive problems in precision medicine.

References

Adamczak, R. and Wolff, P. (2015). Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3):531–586.

Apostol, T. M. et al. (1974). *Mathematical analysis*, volume 2. Addison-Wesley Reading, MA.

Athey, S. and Wager, S. (2017). Efficient policy learning.

Barbe, P. and Bertail, P. (1995). *The weighted bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

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

Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72–e72.

Berkelaar, M. and others (2015). *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*. R package version 5.6.13.

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

Boulesteix, A.-L., Strobl, C., Augustin, T., and Daumer, M. (2008). Evaluating microarray-based classifiers: an overview. *Cancer informatics*, 6:CIN–S408.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723.

Chakraborty, B., Laber, E. B., and Zhao, Y.-Q. (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer Science & Business Media.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Charbonnel, B. H., Matthews, D. R., Schernthaner, G., Hanefeld, M., Brunetti, P., and on behalf of the QUARTET Study Group (2005). A long-term comparison of pioglita-

zone and gliclazide in patients with type 2 diabetes mellitus: a randomized, double-blind, parallel-group comparison trial. *Diabetic Medicine*, 22(4):399–405.

Chen, Y., Liu, Y., Zeng, D., and Wang, Y. (2019). *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*. R package version 1.0.

Cheng, G. and Huang, J. Z. (2010). Bootstrap consistency for general semiparametric m -estimation. *Annals of Statistics*, 38(5):2884–2915.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819.

Cui, Y., Zhu, R., and Kosorok, M. (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic Journal of Statistics*, 11(2):3927–3953.

Datta, A. and Zou, H. (2017). Cocolasso for high-dimensional error-in-variables regression. *Annals of Statistics*.

Dean, II, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., and et al. (2016). Scaling up scientific discovery in sleep medicine: The national sleep research resource. *Sleep*, 39(5):1151–1164.

Díaz, I., Savenkov, O., and Ballman, K. (2018). Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika*, 105(3):723–738.

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018). I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics*, 46(2):814–841.

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

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41(2):361–372.

Giné, E. and Sang, H. (2010). Uniform asymptotics for kernel density estimators with variable bandwidths. *Journal of Nonparametric Statistics*, 22(6):773–795.

Goldberg, Y., Song, R., Zeng, D., and Kosorok, M. R. (2014). Comment on "dynamic treatment regimes: Technical challenges and applications". *Electronic Journal of Statistics*, 8(1):1290.

Greenwood, D. C., Gilthorpe, M. S., and Cade, J. E. (2006). The impact of imprecisely measured covariates on estimating gene-environment interactions. *BMC medical research methodology*, 6(1):21.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120.

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

Jeng, X. J., Lu, W., Peng, H., et al. (2018). High-dimensional inference for personalized treatment decision. *Electronic Journal of Statistics*, 12(1):2074–2089.

Jiang, B., Song, R., Li, J., and Zeng, D. (2019). Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29:1633–1710.

Kim, J. K. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18:191–219.

Kosorok, M. (2010). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York.

Kosorok, M. R. and Moodie, E. E. (2016). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine.* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Laber, E. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225–1272.

Li, K.-C. et al. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248.

Liao, J., Li, X., Wong, T.-Y., Wang, J. J., Khor, C. C., Tai, E. S., Aung, T., Teo, Y.-Y., and Cheng, C.-Y. (2014). Impact of measurement error on testing genetic association with quantitative traits. *PloS one*, 9(1):e87044.

Lin, Y. (2002). Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2017). Interactive q-learning for probabilities and quantiles. *Journal of the American Statistical Association*, 112:638–649.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Lou, Z., Shao, J., and Yu, M. (2018). Optimal treatment assignment to maximize expected outcome with multiple treatments. *Biometrics*, 74(2):506–516.

Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(1):274.

Marcus, C. L., Moore, R. H., Rosen, C. L., Giordani, B., Garetz, S. L., Taylor, H. G., and et al. (2013). A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 368(25):2366–2376. PMID: 23692173.

Mason, D. M. (2012). Proving consistency of non-standard kernel estimators. *Statistical Inference for Stochastic Processes*, 15:151–176.

McKeague, I. W. and Qian, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *Journal of the American Statistical Association*, 110(512):1422–1433.

Mebane, Jr., W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42(11):1–26.

Mei, S., Bai, Y., and Montanari, A. (2018). The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774.

Moodie, E. E. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Murphy, S. A. (2005b). A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Core discussion papers, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

Orellana, L., Rotnitzky, A. G., and Robins, J. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part ii: Proofs of results. *The International Journal of Biostatistics*, 6:Article 9.

Parhofer, K. G. (2015). Interaction between glucose and lipid metabolism: more than diabetic dyslipidemia. *Diabetes & metabolism journal*, 39(5):353–362.

Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 1–18.

Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical applications in genetics and molecular biology*, 4(1).

Qi, Z., Liu, Y., et al. (2018). D-learning to estimate optimal individual treatment rules. *Electronic Journal of Statistics*, 12(2):3601–3638.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180–1210.

Redline, S., Amin, R., Beebe, D., Chervin, R. D., Garetz, S. L., Giordani, B., and et al. (2011). The childhood adenotonsillectomy trial (chat): Rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. *Sleep*, 34(11):1509–1517.

Robins, J., Hernan, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Robins, J., Orellana, L., and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721.

Robins, J. M. (2004). *Optimal Structural Nested Models for Optimal Sequential Decisions*, pages 189–326. Springer New York, New York, NY.

Rocke, D. M. and Durbin, B. (2001). A model for measurement error for gene expression arrays. *Journal of computational biology*, 8(6):557–569.

Rosenbaum, M. and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *Ann. Statist.*, 38(5):2620–2651.

Rosenbaum, M. and Tsybakov, A. B. (2013). *Improved matrix uncertainty selector*, volume Volume 9 of *Collections*, pages 276–290. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Rosenbaum, M., Tsybakov, A. B., et al. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics.

Royden, H. and Fitzpatrick, P. (2010). *Real Analysis*. Prentice Hall.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58.

Rubin, D. B. (1981). The bayesian bootstrap. *Annals of Statistics*, 9(1):130–134.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, 81:961–962.

Rudelson, M., Zhou, S., et al. (2017). Errors-in-variables models with dependent measurements. *Electronic Journal of Statistics*, 11(1):1699–1797.

Shi, C., Fan, A., Song, R., Lu, W., et al. (2018). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of Statistics*, 46(3):925–957.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, New York.

Song, R., Wang, W., Zeng, D., and Kosorok, M. (2015). Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25:901–920.

Sørensen, Ø., Frigessi, A., and Thoresen, M. (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica*, pages 809–829.

Sørensen, Ø., Hellton, K. H., Frigessi, A., and Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. *Journal of Computational and Graphical Statistics*, (just-accepted).

Tadesse, M. G., Ibrahim, J. G., Gentleman, R., Chiaretti, S., Ritz, J., and Foa, R. (2005). Bayesian error-in-variable survival model for the analysis of genechip arrays. *Biometrics*, 61(2):488–497.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

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

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

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Wainwright, M. J. (2015). Basic tail and concentration bounds.

Wang, L., Zhou, Y., Song, R., and Sherwood, B. (2018). Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254.

Wasserman, L. A. (2014). Stein ' s method and the bootstrap in low and high dimensions : A tutorial.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhang, Y., Laber, E. B., Davidian, M., and Tsiatis, A. A. (2018). Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113(524):1541–1549.

Zhao, P. and Xue, L. (2010). Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. *Journal of Multivariate Analysis*, 101(8):1872–1883.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015a). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598. PMID: 26236062.

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015b). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187.

Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B*, 68(3):549–570.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.

Zhu, W., Zeng, D., and Song, R. (2018). Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, pages 1–14.

Appendix A

Supporting Information for Chapter 2

A.1 Chapter Outline

The chapter is constructed as follows. Appendix A.2 states some regularity conditions and useful lemmas for theorems in Chapter 2. Appendix A.3 states a preliminary lemma which validates the idea of estimating the value function $V(\boldsymbol{\beta})$ by its sample analog. Appendix A.4 presents the proofs of technical lemmas mentioned in Appendix A.2. Appendix A.5 includes all proofs of Theorem 2.1–2.4 in Chapter 2. All other technical lemmas mentioned in Appendix A.5 care proved in Appendix A.6. In Appendix A.7, we study the properties of the smoothed estimator and its bootstrapped version under a moving parameter or local asymptotic framework. Proofs of Theorem A.1 and Theorem A.2 in this section can be found in Appendix A.8. Finally, Appendix A.9 presents the pseudo codes for the proximal algorithm we proposed in Section 2.2.4, and Appendix A.10 provides some additional numerical results.

A.2 Regularity Conditions and Useful Lemmas

We first state some regularity conditions, where (K1)–(K3) are assumptions imposed on $K(\cdot)$, while (A1)–(A6) are assumptions imposed on the data.

(K1) $K(\cdot)$ is twice differentiable, $K(\cdot)$, $K'(\cdot)$ and $K''(\cdot)$ all bounded variation on the real line. Furthermore, $\lim_{\nu\to-\infty} K(\nu) = 0$, $\lim_{\nu\to\infty} K(\nu) = 1$; $\int\{K'(\nu)\}^2 d\nu$ and $\int\{K''(\nu)\}^2 d\nu$ are both finite.

(K2) For some integer $b \geqslant 2$, and any $1 \leqslant i \leqslant b$, $\int |\nu^i K'(\nu)| d\nu < \infty$; $\int_{-\infty}^{\infty} \nu^i K'(\nu) d\nu = 0$ for $1 \leqslant i \leqslant b-1$ and $\int_{-\infty}^{\infty} \nu^b K'(\nu) d\nu = d \neq 0$.

(K3) For any integer i between 0 and b, any $\eta > 0$, and any sequence $\{h_n\}$ converging to 0, $\lim_{n\to\infty} h_n^{i-b} \int_{|h_n v|>\eta} |\nu^i K'(\nu)| d\nu = 0$, and $\lim_{n\to\infty} h_n^{-1} \int_{|h_n \nu|>\eta} |K''(\nu)| d\nu = 0$.

(A1) $\mu(a, \boldsymbol{x})$ is bounded for almost all $\boldsymbol{x}$, and $a = 0, 1$; $Y_a^* - \mu(a, \boldsymbol{x})$, $a = 0, 1$, has a sub-Gaussian distribution for almost every $\boldsymbol{x}$.

(A2) The support of the distribution of $\boldsymbol{x}$ is not contained in any proper linear subspace of $\mathbb{R}^p$. For almost every $\widetilde{\boldsymbol{x}}$, the distribution of x_1 conditional on $\widetilde{\boldsymbol{x}}$ has everywhere a positive density. The components of $\widetilde{\boldsymbol{x}}$ are bounded by M_x.

(A3) Let $S(z, \widetilde{\boldsymbol{x}}) = \mathrm{E}\{Y_1^* - Y_0^* | z, \widetilde{\boldsymbol{x}}\}$, where $z = \boldsymbol{x}^T \boldsymbol{\beta}_0$. For almost every $\widetilde{\boldsymbol{x}}$, $S(0, \widetilde{\boldsymbol{x}}) = 0$. And for every $\epsilon > 0$, $\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\epsilon} \mathrm{E}\{\mathrm{I}(x^T\boldsymbol{\beta} > 0) S(z, \widetilde{\boldsymbol{x}}) f(z|\widetilde{\boldsymbol{x}})\} < \mathrm{E}\{\mathrm{I}(x^T\boldsymbol{\beta}_0 > 0) S(z, \widetilde{\boldsymbol{x}}) f(z|\widetilde{\boldsymbol{x}})\}$.

(A4) Given any integer $0 \leqslant i \leqslant b-1$, for all z in a neighborhood of 0, $f^{(i)}(z|\widetilde{\boldsymbol{x}})$ is a continuous function of z and satisfies $|f^{(i)}(z|\widetilde{\boldsymbol{x}})| < M_f$ for almost every $\widetilde{\boldsymbol{x}}$, where $M_f > 0$ is a constant.

(A5) Let $S^{(i)}(0, \widetilde{\boldsymbol{x}})$, $i = 0, 1, \ldots, b$, denote the ith partial derivative of $S(z, \widetilde{\boldsymbol{x}})$ with respect to z. For $0 \leqslant i \leqslant b$, for all z in a neighborhood of 0, $S^{(i)}(z, \widetilde{\boldsymbol{x}})$ is a continuous function of z and satisfies $|S^{(i)}(z, \widetilde{\boldsymbol{x}})| < M_S$ for almost every $\widetilde{\boldsymbol{x}}$, where $M_s > 0$ is a constant. The matrices $\mathrm{E}\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) S^{(1)}(0, \widetilde{\boldsymbol{x}})\}$ and $-\mathrm{E}\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T (\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0) f(0|\widetilde{\boldsymbol{x}}) S^{(1)}(0, \widetilde{\boldsymbol{x}})\}$ are negative definite.

(A6) The random weights $r_1, ..., r_n$ form a random sample from a distribution of a positive random variable with mean one and variance one. Assume that $r_i - \mathrm{E}(r_i)$ has a sub-Gaussian distribution, $i = 1, \ldots, n$.

Remark A.1

The bounded variation assumption on $K(\cdot)$, $K'(\cdot)$ and $K''(\cdot)$ are relatively weak (Chapter 6, Apostol et al. (1974)). This and other assumptions in (K1)-(K2) are satisfied if $K(\cdot)$ is taken to be the distribution function of standard normal distribution ($b = 2$) or the function in Remark 2.1 ($b = 4$). However, $K(\cdot)$ is not required to be a cumulative distribution function. The bounded variation assumption implies that $K(\cdot)$, $|K'(\cdot)|$ and $|K''(\cdot)|$ are uniformly bounded. Our assumptions on the data are also relatively mild. Condition (A1) imposes mild assumption on the tail distribution of $Y_a^* - \mu(a, \boldsymbol{x})$, $a = 0, 1$, and allows for both normal distribution and many other nonnormal distributions. Condition (A3) is a margin type condition to ensure identification of $\boldsymbol{\beta}_0$. □

Let

$$\begin{aligned} \mathcal{G} &= \big\{A\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0)Y + (1-A)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} \leqslant 0)Y : \boldsymbol{\beta} \in \mathbb{B}\big\}, \\ \mathcal{G}^* &= \Big\{(r-1)\{A\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0) + (1-A)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} \leqslant 0)\}Y : \boldsymbol{\beta} \in \mathbb{B}\Big\}. \end{aligned}$$

It is easy to see $\mathcal{G}$ and $\mathcal{G}^*$ are both Donsker classes of functions. Next, we state a useful lemma concerning the Donsker properties of several other classes of functions that involve a smoothing parameter, as well as four technical lemmas that are useful for the proof of the main theorems and are proved based on the Donsker properties using empirical processes techniques. Their proofs can be found in the online supplementary material.

Lemma A.1

Under (K1), (A1)-(A3), the following six classes of functions are Donsker classes.

$$\begin{aligned}
\mathcal{F} &= \big\{(2A-1)K\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)Y : \boldsymbol{\beta}\in\mathbb{B}, h\in(0,1]\big\},\\
\mathcal{F}^* &= \big\{r(2A-1)K\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)Y : \boldsymbol{\beta}\in\mathbb{B}, h\in(0,1]\big\},\\
\mathcal{H} &= \big\{(2A-1)K'\big(\frac{z+\boldsymbol{\psi}^T\widetilde{\boldsymbol{x}}}{h}\big)\widetilde{\boldsymbol{x}}Y : \boldsymbol{\psi}\in\Psi, h\in(0,1]\big\},\\
\mathcal{H}^* &= \big\{r(2A-1)K'\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)\widetilde{\boldsymbol{x}}Y : \boldsymbol{\beta}\in\mathbb{B}, h\in(0,1]\big\},\\
\mathcal{Q} &= \big\{(2A-1)K''\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TY : \boldsymbol{\beta}\in\mathbb{B}, h\in(0,1]\big\},\\
\mathcal{Q}^* &= \big\{r(2A-1)K''\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TY : \boldsymbol{\beta}\in\mathbb{B}, h\in(0,1]\big\},
\end{aligned}$$

where $\Psi=\{\boldsymbol{\psi}:\boldsymbol{\psi}\in\mathbb{R}^{p-1}, ||\boldsymbol{\psi}||\leqslant\frac{\eta}{2\sqrt{p-1}M_x}\}$, with $||\cdot||$ denoting the l_2 norm. □

Lemma A.2

Let $G_i(\boldsymbol{x}_i,\boldsymbol{\beta},h_n)=(2A_i-1)K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)Y_i-\mathrm{E}\big\{(2A_i-1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta}>0)Y_i\big\}$. Under Assumptions (A1)-(A3) and (K1), $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\big|n^{-1}\sum_{i=1}^n G_i(\boldsymbol{x}_i,\boldsymbol{\beta},h_n)\big|\xrightarrow{p}0$. □

Lemma A.3

For any $\boldsymbol{\theta}\in\mathbb{R}^{p-1}$, let $\boldsymbol{R}_n(\boldsymbol{\theta})=\frac{2}{nh_n^2}\sum_{i=1}^n(2A_i-1)K'\big(\frac{z_i}{h_n}+\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}_i\big)\widetilde{\boldsymbol{x}}_iY_i$. let $\eta>0$ be such that $S^{(1)}(z,\widetilde{\boldsymbol{x}})$, $S^{(2)}(z,\widetilde{\boldsymbol{x}})$, and $f^{(1)}(z|\widetilde{\boldsymbol{x}})$ exist and are uniformly bounded for almost every $\widetilde{\boldsymbol{x}}$ if $|z|\leqslant\eta$. Define $\Theta_n=\big\{\boldsymbol{\theta}:\boldsymbol{\theta}\in\mathbb{R}^{p-1}, h_n||\boldsymbol{\theta}||\leqslant\frac{\eta}{2\sqrt{p-1}M_x}\big\}$. Assume the conditions of Theorem 2.2 are satisfied, then (1) $\sup_{\boldsymbol{\theta}\in\Theta_n}||\boldsymbol{R}_n(\boldsymbol{\theta})-\mathrm{E}\boldsymbol{R}_n(\boldsymbol{\theta})||\xrightarrow{p}0$. (2) There are finite numbers α_1 and α_2 such that for all $\boldsymbol{\theta}\in\Theta_n$, $||\mathrm{E}\boldsymbol{R}_n(\boldsymbol{\theta})-\boldsymbol{Q}\boldsymbol{\theta}||\leqslant o(1)+\alpha_1h_n||\boldsymbol{\theta}||+\alpha_2h_n||\boldsymbol{\theta}||^2$ uniformly over $\boldsymbol{\theta}\in\Theta_n$. □

Lemma A.4

Define $G_i^*(\boldsymbol{x}_i,\boldsymbol{\beta},h_n)=(2A_i-1)r_iK\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)Y_i-\mathrm{E}\big\{r_i(2A_i-1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta}>0)Y_i\big\}$. Under

Assumptions (A1)-(A3) and (K1), $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|n^{-1}\sum_{i=1}^{n}G_i^*(\boldsymbol{x}_i,\boldsymbol{\beta},h_n)\right| = o_{p_{rw}}(1)$, where $o_{p_{rw}}(1)$ denotes a random sequence that converges to zero in probability with respect to the joint distribution of (r,w). □

Lemma A.5

Assume the conditions of Theorem 2.4 are satisfied, then $(nh_n)^{1/2}\{T_n^*(\widehat{\boldsymbol{\beta}}_n;h_n)-T_n^*(\boldsymbol{\beta}_0;h_n)\} = o_{p_r}(1)$, where $T_n^*(\boldsymbol{\beta},h_n)$ is defined as follows:

$$\boldsymbol{T}_n^*(\boldsymbol{\beta};h_n) = \frac{\partial\widetilde{M}_n^*(\boldsymbol{\beta},h_n)}{\partial\widetilde{\boldsymbol{\beta}}} = \frac{2}{n}\sum_{i=1}^{n}r_i(2A_i-1)K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i}{h_n}Y_i. \qquad \square$$

A.3 A Preliminary Lemma

Lemma A.6

$\mathrm{E}\big[\{\mathrm{I}(A=1)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}>0)+\mathrm{I}(A=0)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}\leqslant 0)\}Y\big] = \frac{1}{2}V(\boldsymbol{\beta})$. □

Proof of Lemma A.6 By the iterative expectation formula,

$$\begin{aligned}
&\mathrm{E}\big[\{\mathrm{I}(A=1)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}>0)+\mathrm{I}(A=0)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}\leqslant 0)\}Y\big]\\
=\;&\mathrm{E}_{A,\boldsymbol{x}}\big[\{\mathrm{I}(A=1)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}>0)+\mathrm{I}(A=0)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}\leqslant 0)\}E(Y|A,\boldsymbol{x})\big]\\
=\;&\mathrm{E}_{A,\boldsymbol{x}}\{\mathrm{I}(A=1)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}>0)\mu(1,\boldsymbol{x})+\mathrm{I}(A=0)\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}\leqslant 0)\mu(0,\boldsymbol{x})\}\\
=\;&\frac{1}{2}\mathrm{E}_{\boldsymbol{x}}\{\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}>0)\mu(1,\boldsymbol{x})+\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta}\leqslant 0)\mu(0,\boldsymbol{x})\} = \frac{1}{2}V(\boldsymbol{\beta}).
\end{aligned}$$

□

A.4 Proof of Technical Lemmas in Appendix A.2

Proof of Lemma A.1 We give below the proof for $\mathcal{F}$. Proofs for the other classes of functions are similar. Since $K(\cdot)$ is continuous, and has bounded variation on the real line, by Jordan's Theorem in Section 6.3 in Royden and Fitzpatrick (2010), there exist bounded, nondecreasing, right continuous functions K_1 and K_2 on $\mathbb{R}$ such that $K = K_1 - K_2$. Let $\mathcal{F}_1 = \big\{(2A-1)K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)Y : \boldsymbol{\beta} \in \mathbb{B}, h \in (0,1]\big\}$, and $\mathcal{F}_2 = \big\{(2A-1)K_2\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big)Y : \boldsymbol{\beta} \in \mathbb{B}, h \in (0,1]\big\}$. Furthermore, let $\mathcal{F}_{10} = \{K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big) : \boldsymbol{\beta} \in \mathbb{B} \subset \mathbb{R}^p, h \in (0,1]\big\}$. We will first prove $\mathcal{F}_{10}$ is a VC class by similar techniques as in Giné and Sang (2010) and Mason (2012). It is sufficient to show the collection of all subgraphs $S_0 = \Big\{\big\{(\boldsymbol{x},t) : K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big) < t\big\} : K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big) \in \mathcal{F}_{10}\Big\}$ forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$.

Since $K_1(\cdot)$ is a bounded, nondecreasing function, assume $\lim_{x\to-\infty} K_1(x) = m_1$ and $\lim_{x\to\infty} K_1(x) = m_2$. Note that

$$\Big\{(\boldsymbol{x},t) : K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big) < t\Big\} = \big\{(\boldsymbol{x},t) : -\boldsymbol{x}^T\boldsymbol{\beta} + hK_1^{-1}(t) > 0\big\},$$

where $K_1^{-1}(t) = -\infty$ if $t \leqslant m_1$, is $K_1^{-1}(t)$ for $m_1 < t \leqslant m_2$ and is ∞ if $t > m_2$. Let $\psi_{\boldsymbol{\beta},h}(\boldsymbol{x},t) = -\boldsymbol{x}^T\boldsymbol{\beta} + hK_1^{-1}(t)$, $S_1 = \{(\boldsymbol{x},t) : \boldsymbol{x} \in \mathcal{X}, t \in (m_1, m_2]\}$ and $S_2 = \{(\boldsymbol{x},t) : \boldsymbol{x} \in \mathcal{X}, t > m_2\}$. Then for any $\boldsymbol{\beta} \in \mathbb{B} \subset \mathbb{R}^p$, $h \in \big(0,1\big]$,

$$\Big\{(\boldsymbol{x},t) : K_1\big(\frac{\boldsymbol{x}^T\boldsymbol{\beta}}{h}\big) < t\Big\} = \big\{(\boldsymbol{x},t) : \psi_{\boldsymbol{\beta},h}(\boldsymbol{x},t)I((\boldsymbol{x},t) \in S_1) > 0\big\} \cup S_2.$$

Note that $\psi_{\boldsymbol{\beta},h}(\boldsymbol{x},t)$ is in a finite dimensional space of functions when restricted to S_1. This implies the collection $\big\{(\boldsymbol{x},t) : \psi_{\boldsymbol{\beta},h}(\boldsymbol{x},t)I((\boldsymbol{x},t) \in S_1) > 0\big\}$ is a VC subgraph class (Lemma 2.6.15, van der Vaart and Wellner (1996)). $\{S_2\}$ is obviously VC. Hence, S_0 is also VC, and hence Donsker. As $(2A-1)Y$ is square integrable and does not depend on $(\boldsymbol{\beta}, h)$, $\mathcal{F}_1$ is a Donsker class with a square integrable envelope (Theorem 2.10.6, van der Vaart and Wellner (1996)). Similarly, $\mathcal{F}_2$ is also a Donsker class. Then by the Donsker

presentation property, $\mathcal{F}$ is Donsker. □

Proof of Lemma A.2 The Donsker property of $\mathcal{F}$ implies that as $n \to \infty$,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big| 2n^{-1} \sum_{i=1}^{n} \Big[(2A_i - 1) K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) Y_i - \mathrm{E}\big\{(2A_i - 1) K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) Y_i\big\} \Big] \Big| \xrightarrow{p} 0.$$

It is sufficient to show

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big| \mathrm{E}\Big[2(2A_i - 1)\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} Y_i \Big] \Big| \to 0.$$

Note that

$$\begin{aligned} & \mathrm{E}\Big[2(2A_i - 1)\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} Y_i \Big] \\ = \; & \mathrm{E}\Big[\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} \big\{ Y_i^*(1) - Y_i^*(0) \big\} \Big] \\ = \; & \mathrm{E}\Big[\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} \big\{ \mu(1, \boldsymbol{x}_i) - \mu(0, \boldsymbol{x}_i) \big\} \Big]. \end{aligned}$$

According to (A1), we know that $\mu(a, \boldsymbol{x})$ is bounded for almost all $\boldsymbol{x}$, and $a = 0, 1$.

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big| \mathrm{E}\Big[2(2A_i - 1)\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} Y_i \Big] \Big| \leqslant \sup_{\boldsymbol{\beta}\in\mathbb{B}} c\mathrm{E}\Big| K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) \Big|,$$

for some positive constant c. For any positive constant τ,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big| \mathrm{E}\Big[2(2A_i - 1)\big\{ K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\} Y_i \Big] \Big| \leqslant I_1 + I_2,$$

where

$$\begin{aligned} I_1 &= \sup_{\boldsymbol{\beta}\in\mathbb{B}} c\mathrm{E}\Big|\big\{K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\}\mathrm{I}(|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geqslant \tau)\Big|, \\ I_2 &= \sup_{\boldsymbol{\beta}\in\mathbb{B}} c\mathrm{E}\Big|\big\{K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\big\}\mathrm{I}(|\boldsymbol{x}_i^T\boldsymbol{\beta}| < \tau)\Big|. \end{aligned}$$

By the property of $K(\cdot)$, $\forall\tau > 0$, the expectation can be made arbitrary small, uniformly in $\boldsymbol{\beta}$, for all $n \geqslant n_0$, where n_0 is a positive integer. So $I_1 \to 0$. On the other hand,

$$I_2 \leqslant c\sup_{\boldsymbol{\beta}\in\mathbb{B}} P\big(|\boldsymbol{x}_i^T\boldsymbol{\beta}| < \tau\big) = c\sup_{\boldsymbol{\beta}\in\mathbb{B}} P(-\tau - \widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{\beta}} < x_1 < \tau + \widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{\beta}}) \leqslant c'\tau.$$

As τ is an arbitrary positive constant, $I_2 \to 0$. This proves the lemma. □

Proof of Lemma A.3 (1) Let $k_{ni}(\boldsymbol{\psi}) = (2A_i - 1)K'\big(\frac{z_i+\boldsymbol{\psi}^T\widetilde{\boldsymbol{x}}_i}{h}\big)\widetilde{\boldsymbol{x}}_i Y_i$. It suffices to show that

$$\sup_{\boldsymbol{\psi}\in\Psi}\Big|\Big|(nh_n^2)^{-1}\sum_{i=1}^n\big\{k_{ni}(\boldsymbol{\psi}) - \mathrm{E}k_{ni}(\boldsymbol{\psi})\big\}\Big|\Big| \xrightarrow{p} 0,$$

where $\Psi = \big\{\boldsymbol{\psi} : \boldsymbol{\psi} \in \mathbb{R}^{p-1}, ||\boldsymbol{\psi}|| \leqslant \frac{\eta}{2\sqrt{p-1}M_x}\big\}$.

The Donsker property of $\mathcal{H}$ implies that

$$\sup_{\boldsymbol{\psi}\in\Psi}\sup_{h\in(0,1]}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{k_{ni}(\boldsymbol{\psi}) - \mathrm{E}k_{ni}(\boldsymbol{\psi})\big\}\Big|\Big| = O_p(n^{-1/2}).$$

Then since $h_n \to 0$ and $nh_n^4 \to \infty$, we can derive that

$$\begin{aligned} \sup_{\boldsymbol{\psi}\in\Psi}\Big|\Big|(nh_n^2)^{-1}\sum_{i=1}^n\big\{k_{ni}(\boldsymbol{\psi}) - \mathrm{E}k_{ni}(\boldsymbol{\psi})\big\}\Big|\Big| &\leqslant h_n^{-2}\sup_{\boldsymbol{\psi}\in\Psi}\sup_{h\in(0,1]}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{k_{ni}(\boldsymbol{\psi}) - \mathrm{E}k_{ni}(\boldsymbol{\psi})\big\}\Big|\Big| \\ &\leqslant O_p(n^{-1/2}h_n^{-2}) = o_p(1). \end{aligned}$$

(2) $\mathrm{E}\big\{\boldsymbol{R}_n(\boldsymbol{\theta})\big\} = I_{n1} + I_{n2}$, where

$$I_{n1} = \frac{1}{h_n^2}\int_{|z|\leqslant\eta} K'\big(\frac{z}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}S(z,\widetilde{\boldsymbol{x}})f(z|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}}),$$

and

$$I_{n2} = \frac{1}{h_n^2}\int_{|z|>\eta} K'\big(\frac{z}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}S(z,\widetilde{\boldsymbol{x}})f(z|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}}).$$

From (A4) and (A5), we can say that for some $M > 0$,

$$||I_{n2}|| \leqslant \frac{M}{h_n^2}\int_{|z|>\eta} K'\big(\frac{z}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big)dzdP(\widetilde{\boldsymbol{x}}).$$

Let $\zeta = z/h_n + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}$. Since $h_n||\boldsymbol{\theta}|| \leqslant \frac{\eta}{2\sqrt{p-1}M_x}$ and $||\widetilde{\boldsymbol{x}}|| \leqslant \sqrt{p-1}M_x$ by (A3), then $|z| > \eta$ implies that

$$|\zeta| > \frac{\eta}{2h_n}, \quad \text{and} \quad ||I_{n2}|| \leqslant \frac{M}{h_n}\int_{h_n|\zeta|>\eta/2} K'(\zeta)d\zeta.$$

And from (K3), it converges to 0 as $n \to \infty$. Therefore,

$$\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_n}||I_{n2}|| = 0.$$

When $|z| \leqslant \eta$, then we have:

$$S(z,\widetilde{\boldsymbol{x}})f(z|\widetilde{\boldsymbol{x}}) = S^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})z + \big\{S^{(1)}(0,\widetilde{\boldsymbol{x}})f^{(1)}(\epsilon_2|\widetilde{\boldsymbol{x}}) + S^{(1)}(\epsilon_1,\widetilde{\boldsymbol{x}})f^{(1)}(0|\widetilde{\boldsymbol{x}})\big\}z^2,$$

where ϵ_1 and ϵ_2 are between 0 and z. So $I_{n1} = J_{n1} + J_{n2}$, where

$$\begin{aligned} J_{n1} &= \frac{1}{h_n^2}\int_{|z|\leqslant\eta} K'\big(\frac{z}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}zS^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}}) \\ &= \int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n} K'(\zeta)S^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})\widetilde{\boldsymbol{x}}(\zeta - \widetilde{\boldsymbol{x}}^T\boldsymbol{\theta})d\zeta dP(\widetilde{\boldsymbol{x}}), \end{aligned}$$

and

$$\begin{aligned} J_{n2} &= \frac{1}{h_n^2}\int_{|z|>\eta} K'\big(\frac{z}{h_n}+\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}\big\{S^{(1)}(0,\widetilde{\boldsymbol{x}})f^{(1)}(\epsilon_2|\widetilde{\boldsymbol{x}})+S^{(1)}(\epsilon_1,\widetilde{\boldsymbol{x}})f^{(1)}(0|\widetilde{\boldsymbol{x}})\big\}z^2dzdP(\widetilde{\boldsymbol{x}}) \\ &= h_n\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|>\eta/h_n} K'(\zeta)\widetilde{\boldsymbol{x}}\big\{S^{(1)}(0,\widetilde{\boldsymbol{x}})f^{(1)}(\epsilon_2|\widetilde{\boldsymbol{x}})+S^{(1)}(\epsilon_1,\widetilde{\boldsymbol{x}})f^{(1)}(0|\widetilde{\boldsymbol{x}})\big\}(\zeta-\widetilde{\boldsymbol{x}}^T\boldsymbol{\theta})^2d\zeta dP(\widetilde{\boldsymbol{x}}). \end{aligned}$$

Since $\int \zeta K'(\zeta)d\zeta = 0$ by (K2), and $\big|h_n\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big| \leqslant \eta/2$,

$$\Big|\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n}\zeta K'(\zeta)d\zeta\Big| \leqslant \int_{|\zeta|\leqslant\eta/2h_n}\big|\zeta K'(\zeta)\big|d\zeta.$$

By (K2), $\int_{|\zeta|\leqslant\eta/2h_n}\big|\zeta K'(\zeta)\big|d\zeta$ is bounded uniformly over n and $\boldsymbol{\theta}\in\Theta_n$, and converges to 0. So

$$\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_n}\Big|\Big|\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n}\zeta K'(\zeta)S^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})\widetilde{\boldsymbol{x}}d\zeta dP(\widetilde{\boldsymbol{x}})\Big|\Big| = 0.$$

In addition,

$$\begin{aligned} \Big|\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n}K'(\zeta)d\zeta\Big| &\leqslant \big|h_n\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\big|h_n^{-1}\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n}\big|K'(\zeta)\big|d\zeta \\ &\leqslant \frac{\eta}{2h_n}\int_{|\zeta|\geqslant\eta/(2h_n)}\big|K'(\zeta)\big|d\zeta. \end{aligned}$$

Similarly, we also have:

$$\lim_{n\to\infty}\Big|\Big|\sup_{\boldsymbol{\theta}\in\Theta_n}\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|\leqslant\eta/h_n}\zeta K'(\zeta)S^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^Td\zeta dP(\widetilde{\boldsymbol{x}})-\boldsymbol{\theta}^T\boldsymbol{Q}\Big|\Big| = 0.$$

Then for J_{n2}, there is some finite $M>0$, and α_1, α_2 such that:

$$\big|\big|J_{n2}\big|\big| \leqslant Mh_n\int_{|\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}|>\eta/h_n}\big|K'(\zeta)\big|(\zeta-\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}})^2d\zeta dP(\widetilde{\boldsymbol{x}}) \leqslant o(1)+\alpha_1h_n||\boldsymbol{\theta}||+\alpha_2h_n||\boldsymbol{\theta}||^2.$$

In conclusion, $\big|\big|\mathrm{E}\boldsymbol{R}_n(\boldsymbol{\theta})-\boldsymbol{Q}\boldsymbol{\theta}\big|\big| \leqslant o(1)+\alpha_1h_n||\boldsymbol{\theta}||+\alpha_2h_n||\boldsymbol{\theta}||^2$. □

Proof of Lemma A.4 The Donsker property of $\mathcal{F}^*$ implies that as $n \to \infty$,

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| \frac{2}{n} \sum_{i=1}^{n} \Big[(2A_i - 1) r_i K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}}{h_n}\big) Y_i - \mathrm{E}_w \mathrm{E}_{r|w} \big\{ (2A_i - 1) r_i K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}}{h_n}\big) Y_i \big\} \Big] \right| = o_{p_{rw}}(1).$$

It is sufficient to show

$$\begin{aligned} & \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| \mathrm{E}_w \mathrm{E}_{r|w} \Big[2(2A_i - 1) r_i \big\{ K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T \boldsymbol{\beta} > 0) \big\} Y_i \Big] \right| \\ = & \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| \mathrm{E}_w \Big[2(2A_i - 1) \big\{ K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}}{h_n}\big) - \mathrm{I}(\boldsymbol{x}_i^T \boldsymbol{\beta} > 0) \big\} Y_i \Big] \right| \to 0, \end{aligned}$$

which is verified in Lemma A.2. Hence the lemma is proved. □

Proof of Lemma A.5 It follows from Lemma 3 of Cheng and Huang (2010) that it is sufficient to prove

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0|| \leqslant C(nh_n)^{-1/2}} \sqrt{nh_n} \big| T_n^*(\boldsymbol{\beta}; h_n) - T_n^*(\boldsymbol{\beta}_0; h_n) \big| = o_{p_{rw}}(1).$$

According to Lemma 9.14 in Kosorok (2010), $\mathcal{H}^*$ is a bounded uniform entropy integral (BUEI) class, and the proof of Lemma 9.13 implies that $\forall\ 0 < \epsilon < 1$, the ϵ-covering number of $\mathcal{H}^*$ satisfies $N\big(\epsilon||F||, \mathcal{H}^*, L(P)\big) \leqslant \big(\frac{A}{\epsilon}\big)^v$, for some positive constants A and v, and an envelop F. Consider the stochastic process

$$f \mapsto n^{-1/2} \sum_{i=1}^{n} f(W_i, r_i), \quad f \in \mathcal{H}^*, \quad W_i = (A_i, \boldsymbol{x}_i, Y_i).$$

Given $Y, \boldsymbol{x}, r$ all have sub-Gaussian distributions, $(f - \mathbb{P}f)$ is a separable sub-Gaussian process. Since $K''(\cdot)$ is bounded, we can derive by the Lipschitz property of $K'(\cdot)$ that

$$||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leqslant C(nh_n)^{-1/2} \implies ||f - f_0|| \leqslant C'(nh_n^3)^{-1/2}.$$

By the property of the increments for the separable sub-Gaussian process (Corollary 2.2.8 in van der Vaart and Wellner (1996)),

$$\begin{aligned}
\mathrm{E}_w\mathrm{E}_{r|w} \quad & \Big[\quad \sup_{||f-f'||\leqslant C'(nh_n^3)^{-1/2}} \Big|\Big| n^{-1/2} \sum_{i=1}^{n} \big[f(W_i, r_i) - \mathrm{E}_w\mathrm{E}_{r|w} f \big] \\
& - \quad n^{-1/2} \sum_{i=1}^{n} \big\{ f'(W_i, r_i) - \mathrm{E}_w\mathrm{E}_{r|w} f' \big\} \Big|\Big| \Big] \\
\leqslant \quad & D \quad \int_0^{C'(nh_n^3)^{-1/2}} \sqrt{\log(A/\epsilon)^v} d\epsilon \leqslant D' \big(nh_n^3\big)^{-1/2} \sqrt{\frac{1}{2}\log\big(nh_n^3\big)},
\end{aligned}$$

for some positive constants D and D'. Then by Markov inequlity, for any $\delta > 0$,

$$\begin{aligned}
& P_{r,w}\Big(\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leqslant C(nh_n)^{-1/2}} \sqrt{nh_n} \big|\big| T_n^*(\boldsymbol{\beta}; h_n) - T_n^*(\boldsymbol{\beta}_0; h_n) \big|\big| > \delta \Big) \\
\leqslant \quad & P_{r,w}\Big(\sup_{||f-f_0||\leqslant C'(nh_n^3)^{-1/2}} h_n^{-1/2} \big|\big| n^{-1/2} \big(\sum_{i=1}^{n} f(W_i, r_i) - \sum_{i=1}^{n} f_0(W_i, r_i) \big) \big|\big| > \delta \Big) \\
\leqslant \quad & (\delta h_n^{1/2})^{-1} \mathrm{E}_w\mathrm{E}_{r|w} \Big[\sup_{||f-f_0||\leqslant C'(nh_n^3)^{-1/2}} \big|\big| n^{-1/2} \big\{ \sum_{i=1}^{n} f(w_i, r_i) - \sum_{i=1}^{n} f_0(W_i, r_i) \big\} \big|\big| \Big] \\
\leqslant \quad & (\delta h_n^{1/2})^{-1} \mathrm{E}_w\mathrm{E}_{r|w} \Big[\sup_{||f-f_0||\leqslant C'(nh_n^3)^{-1/2}} n^{-1/2} \Big|\Big| \sum_{i=1}^{n} \big\{ f(W_i, r_i) - \mathrm{E}_w\mathrm{E}_{r|w} f \big\} \\
& - \sum_{i=1}^{n} \big\{ f_0(W_i, r_i) - \mathrm{E}_w\mathrm{E}_{r|w} f_0 \big\} \Big|\Big| \Big] + (\delta h_n^{1/2})^{-1} \sup_{||f-f_0||\leqslant C'(nh_n^3)^{-1/2}} \big|\big| n^{-1/2} \mathrm{E}_w\mathrm{E}_{r|w} (f - f_0) \big|\big| \\
\leqslant \quad & D'\delta^{-1}(nh_n^4)^{-1/2} \sqrt{\frac{1}{2}\log\big(nh_n^3\big)} + C'\delta^{-1}(nh_n^2)^{-1} \to 0,
\end{aligned}$$

given $\log(n)/nh_n^4 = o_p(1)$, where $P_{r,w}(\cdot)$ denotes probability with respect to the joint distribution of (r, w). The conclusion follows as $\delta > 0$ is arbitrary. □

A.5 Proof of Theorems 2.1–2.4

Proof of Theorem 2.1 We observe that $\widehat{\boldsymbol{\beta}}_n$ maximizes $\widetilde{M}_n(\boldsymbol{\beta}, h_n)$ over $\boldsymbol{\beta} \in \mathbb{B}$. Lemma A.2

implies that $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\widetilde{M}_n(\boldsymbol{\beta},h_n) - M(\boldsymbol{\beta})\right| \to 0$ in probability as $n \to \infty$, where $M(\boldsymbol{\beta}) = \mathrm{E}\{(2A_i-1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)Y_i\}$. Condition (A3) implies that for every $\epsilon > 0$, $\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\epsilon} M(\boldsymbol{\beta}) < M(\boldsymbol{\beta}_0)$. Hence, $\widehat{\boldsymbol{\beta}}_n$ is consistent by Theorem 5.7 in van der Vaart (2000). □

The asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_n$ depends critically on the properties of the gradient and the Hessian matrix of the objective function. $\widetilde{M}_n(\boldsymbol{\beta},h_n)$. Define

$$\boldsymbol{T}_n(\boldsymbol{\beta};h_n) = \frac{\partial\widetilde{M}_n(\boldsymbol{\beta},h_n)}{\partial\widetilde{\boldsymbol{\beta}}} = \frac{2}{n}\sum_{i=1}^{n}(2A_i-1)K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i}{h_n}Y_i, \tag{A.1}$$

$$\boldsymbol{Q}_n(\boldsymbol{\beta};h_n) = \frac{\partial^2\widetilde{M}_n(\boldsymbol{\beta},h_n)}{\partial\widetilde{\boldsymbol{\beta}}\partial\widetilde{\boldsymbol{\beta}}^T} = \frac{2}{n}\sum_{i=1}^{n}(2A_i-1)K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n^2}Y_i. \tag{A.2}$$

Lemmas A.7 and A.8 below establish useful properties of $\boldsymbol{T}_n(\boldsymbol{\beta};h_n)$ and $\boldsymbol{Q}_n(\boldsymbol{\beta};h_n)$, respectively. The proofs of these two lemmas are given in Appendix A.6.

Lemma A.7

Assume the conditions of Theorem 2.2 are satisfied, then

$$\lim_{n\to\infty}\mathrm{E}\{(nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\} = 0, \quad \text{and} \quad \lim_{n\to\infty}\mathrm{Var}\{(nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\} = \boldsymbol{D}. \qquad \square$$

Lemma A.8

Let $\boldsymbol{\beta}_n^r$ be any value between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Assume the conditions of Theorem 2.2 are satisfied, then $\boldsymbol{Q}_n(\boldsymbol{\beta}_n^r;h_n) \xrightarrow{p} \boldsymbol{Q}$, where $\boldsymbol{Q}$ is defined in (2.12). □

Let $V''(\cdot)$ be the Hessian matrix of $V(\cdot)$ with respect to $\widetilde{\boldsymbol{\beta}}$, i.e., $\frac{\partial^2 V(\boldsymbol{\beta})}{\partial\widetilde{\boldsymbol{\beta}}\partial\widetilde{\boldsymbol{\beta}}^T}$. Lemmas A.9 below describes the continuity of $V''(\cdot)$. The proof is given in Appendix A.6.

Lemma A.9

Let $\boldsymbol{\beta}_n^r$ be any value between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. Assume the conditions of Theorem 2.2 are satisfied, then $V''(\boldsymbol{\beta}_n^r) \xrightarrow{p} \boldsymbol{I}_V$, where $\boldsymbol{I}_V = -\mathrm{E}\big\{S^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0)\big\}$. □

Proof of Theorem 2.2 By Taylor expansion, we have

$$\boldsymbol{T}_n(\widehat{\boldsymbol{\beta}}_n; h_n) = \boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n) + \boldsymbol{Q}_n(\boldsymbol{\beta}_n^r; h_n)(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0),$$

where $\boldsymbol{\beta}_n^r$ is between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$. The definition of $\widehat{\boldsymbol{\beta}}_n$ implies that $\boldsymbol{T}_n(\widehat{\boldsymbol{\beta}}_n; h_n) = \widetilde{\mathbf{0}}$, where $\widetilde{\mathbf{0}}$ denotes a $(p-1)$-dimensional vector of zeroes. Lemma A.8 indicates that

$$\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0 = \big(-\boldsymbol{Q} + o_p(1)\big)^{-1} \boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n).$$

To prove the theorem, it suffices to verify $(nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n) \xrightarrow{d} N(0, \boldsymbol{D})$. It is known from Lemma B1 that $\mathrm{E}(nh_n)^{1/2}\big\{\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n)\big\} \to 0$. It is sufficient to prove that $(nh_n)^{1/2}\boldsymbol{\gamma}^T\big\{\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n) - \mathrm{E}\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n)\big\}$ is asymptotically $N(\widetilde{\mathbf{0}}, \boldsymbol{\gamma}^T\boldsymbol{D}\boldsymbol{\gamma})$ for any constant vector $\boldsymbol{\gamma} \in \mathbb{R}^{p-1}$ such that $||\boldsymbol{\gamma}|| = 1$. Let

$$q_i = 2(2A_i - 1)(nh_n)^{1/2} K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i}{h_n} Y_i.$$

It follows the proof of Lemma A.7 that $\lim_{n\to\infty} \mathrm{E}q_i = 0$, and $\lim_{n\to\infty} \mathrm{E}q_i^2/n = \boldsymbol{\gamma}^T\boldsymbol{D}\boldsymbol{\gamma}$.

To apply Lyapunov central limit theorem, we will verify that

$$\lim_{n\to\infty} (s_n^4)^{-1} \sum_{i=1}^{n} \mathrm{E}\big\{(q_i - \mathrm{E}q_i)^4\big\} = 0, \tag{A.3}$$

where $\lim_{n\to\infty}(n^{-2}s_n^2) = \lim_{n\to\infty}\sum_{i=1}^n \mathrm{Var}(n^{-1}q_i) = \boldsymbol{\gamma}^T\boldsymbol{D}\boldsymbol{\gamma}$. We observe that the left-side of (A.3) is bounded from above (up to a positive constant) by $\lim_{n\to\infty} n^{-3}\mathrm{E}(q_i^4) + \lim_{n\to\infty} n^{-3}(\mathrm{E}q_i)^4 = I_1 + I_2$. As $(\mathrm{E}q_i)^4 \to 0$, we have $I_2 = o(1)$. To evaluate I_1, note that

$$n^{-3}\mathrm{E}(q_i^4) = 16(nh_n^2)^{-1}\mathrm{E}\big\{K'\big(\frac{x_i^T\boldsymbol{\beta}}{h_n}\big)^4(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i)^4 Y_i^4\big\}.$$

Since Y has sub-Gaussian tail, then for any integer $k \geqslant 1$, $\mathrm{E}|Y|^k$ is finite. So with the boundedness of $K(\cdot)$ and $\widetilde{\boldsymbol{x}}$, $\mathrm{E}\big\{K'\big(\frac{x_i^T\boldsymbol{\beta}}{h_n}\big)^4(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i)^4 Y_i^4\big\}$ is finite. Then $n^{-1}h_n^{-2} = o(1)$ implies $I_1 = o(1)$. Therefore, the Lyapunov condition is satisfied. This proves (1).

To prove (2), we observe that

$$\begin{aligned}\sqrt{n}(V_n(\widehat{\boldsymbol{\beta}}_n) - V(\boldsymbol{\beta}_0)) &= \sqrt{n}\big\{V_n(\boldsymbol{\beta}_0) - V(\boldsymbol{\beta}_0)\big\} + \sqrt{n}\big\{V_n(\widehat{\boldsymbol{\beta}}_n) - V_n(\boldsymbol{\beta}_0) - V(\widehat{\boldsymbol{\beta}}_n) + V(\boldsymbol{\beta}_0)\big\} \\ &\quad + \sqrt{n}\big\{V(\widehat{\boldsymbol{\beta}}_n) - V(\boldsymbol{\beta}_0)\big\} \\ &= I_1 + I_2 + I_3,\end{aligned}$$

where the definition of I_i ($i = 1, 2, 3$) is clear from the context. By the central limit theorem, we have $I_1 \xrightarrow{d} N(0, U)$. The Donsker property of the function class $\mathcal{G}$ ensures that $I_2 = o_p(1)$. Note that with the consistency result in Theorem 2.1, we have $P(\widehat{\boldsymbol{\beta}}_{n1} = \boldsymbol{\beta}_{01}) \to 1$ as $n \to \infty$. By Taylor expansion, we have

$$I_3 = \sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0)^T V'(\boldsymbol{\beta}_0) + \sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0)^T V''(\boldsymbol{\beta}^r)(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)/2 + o_p(1),$$

where $V'(\cdot)$ and $V''(\cdot)$ denote the gradient vector and Hessian matrix of $V(\cdot)$ with respect to $\widetilde{\boldsymbol{\beta}}$, respectively; $\boldsymbol{\beta}^r$ is between $\widetilde{\boldsymbol{\beta}}_0$ and $\widetilde{\boldsymbol{\beta}}_n$. As $\boldsymbol{\beta}_0$ is the maximizer of $V(\cdot)$, we have $V'(\boldsymbol{\beta}_0) = 0$. Let $\lambda_{max}(\cdot)$ be the eigenvalue with the greatest absolute value. The second term is upper bounded by $|\lambda_{max}(V''(\boldsymbol{\beta}^r))|\sqrt{n}||\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0||^2/2$, which is of order $O_p(n^{-1/2}h^{-1}) = o_p(1)$ by Lemma A.9, Assumption (A5) and the first part of the theorem on the convergence rate. This proves (2). □

In the rest of this appendix, we will prove the theory for bootstrap based inference. As described in Section 2.3.2, given a sequence of random variables R_n, $n = 1, \ldots, n$, we write $R_n = o_{p_r}(1)$ if for any $\epsilon > 0, \delta > 0$, we have $P_w\big(P_{r|w}(|R_n| > \epsilon) > \delta\big) \to 0$ as $n \to \infty$. Furthermore, $o_{p_{rw}}(1)$ denotes a random sequence that converges to zero in probability with respect to the joint distribution of (r, w); and $o^*_{P_W}(1)$ denotes a random sequence that converges to zero in probability with respect to the distribution of r only. By Lemma 3 of Cheng and Huang (2010), if $R_n = o_{p_{rw}}(1)$, then $R_n = o_{p_r}(1)$. In particular, if R_n depends only on the data w but not on the random weights r and if $R_n = o_{p_w}(1)$, then

it is easy to see $R_n = o_{p_{rw}}(1)$, and hence it is $o_{p_r}(1)$. In this part of proof, we will include subscripts in the probability and expectation to clarify which probability distribution is used in the calculation.

Proof of Theorem 2.3 By definition, $\widehat{\boldsymbol{\beta}}_n^*$ maximizes $\widetilde{M}_n^*(\boldsymbol{\beta}, h_n)$ over $\boldsymbol{\beta} \in \mathbb{B}$. First, by combining Lemma A.2 and Lemma A.4 and recognizing that $\mathrm{E}_w\{(2A_i - 1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)Y_i\} = \mathrm{E}_w\mathrm{E}_{r|w}\big\{r_i(2A_i - 1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)Y_i\big\}$, we have $\sup_{\boldsymbol{\beta}\in\mathbb{B}} |\widetilde{M}_n^*(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\boldsymbol{\beta}, h_n)| = o_{p_{rw}}(1)$. By Lemma 3 of Cheng and Huang (2010), $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\widetilde{M}_n^*(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\boldsymbol{\beta}, h_n)\big| = o_{p_r}(1)$. By Theorem 5.7 in van der Vaart (2000), to prove the theorem, it is sufficient to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P_w\Big(\sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon} \big\{\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\widehat{\boldsymbol{\beta}}_n, h_n)\big\} < 0\Big) = 1. \tag{A.4}$$

By Lemma A2, $\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\widehat{\boldsymbol{\beta}}_n, h_n) = M(\boldsymbol{\beta}) - M(\widehat{\boldsymbol{\beta}}_n) + o_{p_w}(1)$. Furthermore, the consistency of $\widehat{\boldsymbol{\beta}}_n$ implies that for all sufficiently large n, any $\boldsymbol{\beta}$ that satisfies $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n|| > \epsilon$ would also satisfy $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \geqslant \epsilon/2$. Condition (A3) implies that $\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\epsilon/2} M(\boldsymbol{\beta}) < M(\boldsymbol{\beta}_0)$. Hence, (A.4) holds. This proves (1).

To prove (2), we observe that $\sqrt{n}(V_n^*(\boldsymbol{\beta}) - V_n(\boldsymbol{\beta})) = n^{-1/2}\sum_{i=1}^n (r_i - 1)\big\{A_i\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) + (1 - A_i)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} \leqslant 0)\big\}Y_i$, which has mean zero. The Donsker property of the function class $\mathcal{G}^*$ and the fact $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + o_{p_{rw}}(1)$ implies that

$$\sqrt{n}\big[\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\} - \{V_n^*(\boldsymbol{\beta}_0) - V_n(\boldsymbol{\beta}_0)\}\big] = o_{p_{rw}}(1), \tag{A.5}$$

by Lemma 19.24 of van der Vaart (2000). By assumption (A6) and the classical central limit theorem, $\sqrt{n}\{V_n^*(\boldsymbol{\beta}_0) - V_n(\boldsymbol{\beta}_0)\} = N(0, U) + o_{p_{rw}}(1)$. Hence, $\sqrt{n}\big\{V_n^*(\widehat{\boldsymbol{\beta}}_n) - V_n(\widehat{\boldsymbol{\beta}}_n)\big\} = N(0, U) + o_{p_{rw}}(1)$. Lemma 3 in Cheng and Huang (2010) implies (2) holds. □

To prove Theorem 2.4, we define the following gradient function and Hessian matrix

corresponding to the randomly weighted objective function

$$\boldsymbol{T}_n^*(\boldsymbol{\beta}; h_n) = \frac{\partial \widetilde{M}_n^*(\boldsymbol{\beta}, h_n)}{\partial \widetilde{\boldsymbol{\beta}}} = \frac{2}{n}\sum_{i=1}^{n} r_i(2A_i - 1)K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i}{h_n}Y_i, \tag{A.6}$$

$$\boldsymbol{Q}_n^*(\boldsymbol{\beta}; h_n) = \frac{\partial^2 \widetilde{M}_n^*(\boldsymbol{\beta}, h_n)}{\partial \widetilde{\boldsymbol{\beta}} \partial \widetilde{\boldsymbol{\beta}}^T} = \frac{2}{n}\sum_{i=1}^{n} r_i(2A_i - 1)K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n^2}Y_i. \tag{A.7}$$

Lemma A.10 below characterizes the asymptotic property of the Hessian matrix. Its proof is given in the supplementary material.

Lemma A.10

Let $\boldsymbol{\beta}_n^{*r}$ be a variable between $\widehat{\boldsymbol{\beta}}_n^*$ and $\widehat{\boldsymbol{\beta}}_n$. Assume the conditions of Theorem 2.4 are satisfied, then $\boldsymbol{Q}_n^*(\boldsymbol{\beta}_n^{*r}; h_n) = \boldsymbol{Q} + o_{p_r}(1)$. □

Proof of Theorem 2.4 By Taylor expansion, $\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n^*; h_n) = \boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n; h_n) + \boldsymbol{Q}_n^*(\boldsymbol{\beta}_n^{*r}; h_n)(\widetilde{\boldsymbol{\beta}}_n^* - \widetilde{\boldsymbol{\beta}}_n)$, where $\boldsymbol{\beta}_n^{*r}$ is between $\widehat{\boldsymbol{\beta}}_n^*$ and $\widehat{\boldsymbol{\beta}}_n$. By the definition of $\widehat{\boldsymbol{\beta}}_n^*$, we have $\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n^*; h_n) = \widetilde{\boldsymbol{0}}$. By Lemma B4, we have

$$\widetilde{\boldsymbol{\beta}}_n^* - \widetilde{\boldsymbol{\beta}}_n = -\big(\boldsymbol{Q} + o_{p_r}(1)\big)^{-1}\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n; h_n).$$

It remains to show $(nh_n)^{1/2}\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n; h_n) = N(0, \boldsymbol{D}) + o_{p_r}(1)$. By Lemma A5, we only need to show $(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n) = N(0, \boldsymbol{D}) + o_{p_r}(1)$. Observe that

$$\begin{aligned}
\mathrm{E}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} &= (nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n), \\
\mathrm{Var}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} &= \frac{4}{n}\sum_{i=1}^{n}\big\{K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)\big\}^2\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n}Y_i^2.
\end{aligned}$$

Lemma A.7 implies that

$$\lim_{n\to\infty}\mathrm{E}_w\mathrm{E}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} = 0, \quad \text{and} \quad \lim_{n\to\infty}\mathrm{E}_w\Big[\mathrm{Var}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\}\Big] = \boldsymbol{D}.$$

It suffices to prove that for any constant vector $\boldsymbol{\gamma} \in \mathbb{R}^{p-1}$ such that $||\boldsymbol{\gamma}|| = 1$,

$$(nh_n)^{1/2}\boldsymbol{\gamma}^T\big\{\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n) - \mathrm{E}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} = N(\widetilde{\mathbf{0}}, \boldsymbol{\gamma}^T\boldsymbol{D}\boldsymbol{\gamma}) + o_{p_r}(1).$$

Define $q_i^* = 2r_i(2A_i - 1)(nh_n)^{1/2}K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i}{h_n}Y_i$, where $\mathrm{E}_{r|w}q_i^* = q_i$, and $\mathrm{E}_{r|w}(q_i^{*2}) = 2q_i^2$, for q_i defined in the proof of Theorem 2.2. To check the Lyapunov condition, it suffices to prove that

$$(s_n^{*4})^{-1}\sum_{i=1}^n \mathrm{E}_{r|w}\big\{(q_i^* - \mathrm{E}_{r|w}q_i^*)^4\big\} \xrightarrow{a.s.} 0,$$

where $s_n^{*2} = \sum_{i=1}^n \mathrm{Var}_{r|w}(q_i^*)$. Similarly as Theorem 2, the Lyapunov condition holds if

$$(s_n^{*4})^{-1}\sum_{i=1}^n \mathrm{E}_{r|w}(q_i^{*4}) \xrightarrow{a.s.} 0, \quad \text{and} \quad (s_n^{*4})^{-1}\sum_{i=1}^n (\mathrm{E}_{r|w}q_i^*)^4 \xrightarrow{a.s.} 0.$$

Since r and Y both have sub-Gaussian tails, we know that for any integer $k \geqslant 1$, $\mathrm{E}|r|^k$ and $\mathrm{E}|Y|^k$ are finite. Then it is easy to compute that $s_n^{*2} = 4n^2h_n^{-1}I_1$, $\sum_{i=1}^n \mathrm{E}_{r|w}(q_i^{*4}) = 16n^3h_n^{-2}\mathrm{E}(r^4)I_2$, and $\sum_{i=1}^n(\mathrm{E}_{r|w}q_i^*)^4 = 16n^3h_n^{-2}I_2$, where

$$I_1 = n^{-1}\sum_{i=1}^n \big\{K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)^2(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i)^2Y_i^2\big\}, \qquad I_2 = n^{-1}\sum_{i=1}^n \big\{K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)^4(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i)^4Y_i^4\big\}.$$

According to (K1) and (A1)-(A2), we know that I_1 and I_2 are both absolutely integrable. Then the strong law of large numbers implies that $I_1 \xrightarrow{a.s.} \mathrm{E}_wI_1$ and $I_2 \xrightarrow{a.s.} \mathrm{E}_wI_2$. With the continuous mapping theorem, it is easy to conclude that $I_1^{-2}I_2 \xrightarrow{a.s.} (\mathrm{E}_wI_1)^{-2}\mathrm{E}_wI_2$. We therefore have

$$(s_n^{*4})^{-1}\sum_{i=1}^n \mathrm{E}_{r|w}(q_i^{*4}) = n^{-1}\mathrm{E}(r_i^4)I_1^{-2}I_2 \xrightarrow{a.s.} 0, \quad (s_n^{*4})^{-1}\sum_{i=1}^n(\mathrm{E}_{r|w}q_i^*)^4 = n^{-1}I_1^{-2}I_2 \xrightarrow{a.s.} 0.$$

This verifies the Lyapunov condition and finishes the proof. □

A.6 Proof of Auxiliary Results in Appendix A.5

Proof of Lemma A.7 (1) Let $\zeta = z/h_n$, then by (A1), we have

$$\begin{aligned}
\mathrm{E}\big\{h_n^{-b}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\} &= h_n^{-b}\mathrm{E}\Big\{(2A-1)K'\big(\frac{x^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}}{h_n}Y\Big\} \\
&= h_n^{-b}\mathrm{E}\Big\{K'\big(\frac{x^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}}{h_n}(Y_1^*-Y_0^*)\Big\} \\
&= h_n^{-b}\mathrm{E}\Big\{K'\big(\frac{z}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}}{h_n}S(z,\widetilde{\boldsymbol{x}})\Big\} \\
&= h_n^{-b}\int K'(\zeta)\widetilde{\boldsymbol{x}}S(h_n\zeta,\widetilde{\boldsymbol{x}})f(h_n\zeta|\widetilde{\boldsymbol{x}})d\zeta dP(\widetilde{\boldsymbol{x}}).
\end{aligned}$$

Under (A3), $S(0,\widetilde{\boldsymbol{x}}) = 0$ for almost every $\widetilde{\boldsymbol{x}}$, so the Taylor series expansions for $S(h_n\zeta,\widetilde{\boldsymbol{x}})$ and $f(h_n\zeta|\widetilde{\boldsymbol{x}})$ can be written as

$$\begin{aligned}
S(h_n\zeta,\widetilde{\boldsymbol{x}}) &= \sum_{i=1}^{b-1}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^i}{i!} + S^{(b)}(\xi_b,\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^b}{b!}, \\
f(h_n\zeta|\widetilde{\boldsymbol{x}}) &= \sum_{j=0}^{b-i-1}f^{(j)}(0|\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^j}{(j)!} + f^{(b-i)}(\xi_i|\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^{b-i}}{(b-i)!},
\end{aligned}$$

for $i = 1,...,b-1$, where $\xi_1,...,\xi_b$ are scalars with values between 0 and $h_n\zeta$. Combining these two expansions yields

$$\begin{aligned}
S(h_n\zeta,\widetilde{\boldsymbol{x}})f(h_n\zeta|\widetilde{\boldsymbol{x}}) = {} & S^{(b)}(\xi_b,\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^b}{b!}f(h_n\zeta|\widetilde{\boldsymbol{x}}) + \sum_{i=1}^{b-1}S^{(i)}(0,\widetilde{\boldsymbol{x}})f^{(b-i)}(\xi_i|\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^b}{i!(b-i)!} \\
& + \sum_{i=1}^{b-1}\sum_{j=0}^{b-i-1}S^{(i)}(0,\widetilde{\boldsymbol{x}})f^{(j)}(0|\widetilde{\boldsymbol{x}})\frac{(h_n\zeta)^{i+j}}{i!j!}.
\end{aligned} \tag{A.8}$$

So we have

$$\begin{aligned}
\mathrm{E}\big\{h_n^{-b}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\} &= \int \zeta^b K'(\zeta)\widetilde{\boldsymbol{x}}\Big\{S^{(b)}(\xi_b,\widetilde{\boldsymbol{x}})\frac{f(h_n\zeta|\widetilde{\boldsymbol{x}})}{b!}+\sum_{i=1}^{b-1}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{f^{(b-i)}(\xi_i|\widetilde{\boldsymbol{x}})}{i!(b-i)!}\Big\}d\zeta dP(\widetilde{\boldsymbol{x}})\\
&\quad+\sum_{i=1}^{b-1}\sum_{j=0}^{b-i-1}\int h_n^{i+j-b}\zeta^{i+j}K'(\zeta)\widetilde{\boldsymbol{x}}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{f^{(j)}(0|\widetilde{\boldsymbol{x}})}{i!j!}d\zeta dP(\widetilde{\boldsymbol{x}})\\
&= I_1+I_2+I_3,
\end{aligned}$$

where for some $\eta>0$,

$$\begin{aligned}
I_1 &= \int \zeta^b K'(\zeta)\widetilde{\boldsymbol{x}}\Big\{S^{(b)}(\xi_b,\widetilde{\boldsymbol{x}})\frac{f(h_n\zeta|\widetilde{\boldsymbol{x}})}{b!}+\sum_{i=1}^{b-1}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{f^{(b-i)}(\xi_i|\widetilde{\boldsymbol{x}})}{i!(b-i)!}\Big\}d\zeta dP(\widetilde{\boldsymbol{x}}),\\
I_2 &= \sum_{i=1}^{b-1}\sum_{j=0}^{b-i-1}\int_{|h_n\zeta|\leqslant\eta} h_n^{i+j-b}\zeta^{i+j}K'(\zeta)\widetilde{\boldsymbol{x}}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{f^{(j)}(0|\widetilde{\boldsymbol{x}})}{i!j!}d\zeta dP(\widetilde{\boldsymbol{x}}),\\
I_3 &= \sum_{i=1}^{b-1}\sum_{j=0}^{b-i-1}\int_{|h_n\zeta|>\eta} h_n^{i+j-b}\zeta^{i+j}K'(\zeta)\widetilde{\boldsymbol{x}}S^{(i)}(0,\widetilde{\boldsymbol{x}})\frac{f^{(j)}(0|\widetilde{\boldsymbol{x}})}{i!j!}d\zeta dP(\widetilde{\boldsymbol{x}}).
\end{aligned}$$

Then from (K2), (A4)-(A5), and Lebesgue's dominated convergence theorem, we have $I_1\to\boldsymbol{H}$, where $\boldsymbol{H}=a_H\sum_{i=1}^{b}\frac{1}{i!(b-i)!}\mathrm{E}\big\{\widetilde{\boldsymbol{x}}S^{(i)}(0,\widetilde{\boldsymbol{x}})f^{(b-i)}(0|\widetilde{\boldsymbol{x}})\big\}$ with $a_H=\int\nu^bK'(\nu)d\nu$. Similarly, let $\eta\to0$, we have $I_2\to0$ and $I_3\to0$. Therefore, $\lim\limits_{n\to\infty}\mathrm{E}\big\{(nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\}=0$, since $nh_n^{2b+1}=o(1)$.

(2) Let $t_i=2(2A_i-1)K'\big(\frac{x_i^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i}{h_n}Y_i$, then we have:

$$\begin{aligned}
\mathrm{Var}\big\{(nh_n)^{1/2}\boldsymbol{T}_n\big\} &= h_n\big(\mathrm{E}t_it_i^T-\mathrm{E}\boldsymbol{T}_n\mathrm{E}\boldsymbol{T}_n^T\big)\\
&= 4\mathrm{E}\Big\{K'\big(\frac{x_i^T\boldsymbol{\beta}_0}{h_n}\big)\Big\}^2\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n}Y_i^2-h_n\mathrm{E}\boldsymbol{T}_n\mathrm{E}\boldsymbol{T}_n^T\\
&= 2\mathrm{E}\Big\{K'\big(\frac{z}{h_n}\big)^2\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T}{h_n}\mathrm{E}\big(Y_1^{*2}+Y_0^{*2}|z,\widetilde{\boldsymbol{x}}\big)\Big\}-h_n\mathrm{E}\boldsymbol{T}_n\mathrm{E}\boldsymbol{T}_n^T\\
&= 2\int K'\big(\frac{z}{h_n}\big)^2\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T}{h_n}\mathrm{E}\big(Y_1^{*2}+Y_0^{*2}|z,\widetilde{\boldsymbol{x}}\big)f(z|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}})-h_n\mathrm{E}\boldsymbol{T}_n\mathrm{E}\boldsymbol{T}_n^T\\
&= 2\int K'(\zeta)^2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T\mathrm{E}\big(Y_1^{*2}+Y_0^{*2}|h_n\zeta,\widetilde{\boldsymbol{x}}\big)f(h_n\zeta|\widetilde{\boldsymbol{x}})d\zeta dP(\widetilde{\boldsymbol{x}})-h_n\mathrm{E}\boldsymbol{T}_n\mathrm{E}\boldsymbol{T}_n^T.
\end{aligned}$$

Since $h_n \to 0$, by (A1)-(A2) and the dominated convergence theorem, we have

$$\begin{aligned}
&\lim_{n\to\infty} \big\{ \mathrm{Var}(nh_n)^{1/2}\boldsymbol{T}_n \big\} \\
=\ &\lim_{n\to\infty} \Big\{ 2\int K'(\zeta)^2 \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T \mathrm{E}\big(Y_1^{*2} + Y_0^{*2} | h_n\zeta, \widetilde{\boldsymbol{x}}\big) f(h_n\zeta|\widetilde{\boldsymbol{x}}) d\zeta dP(\widetilde{\boldsymbol{x}}) - h_n \mathrm{E}\boldsymbol{T}_n \mathrm{E}\boldsymbol{T}_n^T \Big\} \\
=\ &2\int K'(\zeta)^2 d\zeta \int \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T \mathrm{E}\big(Y_1^{*2} + Y_0^{*2} | 0, \widetilde{\boldsymbol{x}}\big) f(0|\widetilde{\boldsymbol{x}}) dP(\widetilde{\boldsymbol{x}}) - 0 * \boldsymbol{H}\boldsymbol{H}^T \\
=\ &a_1 \int \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}}) \mathrm{E}\big(Y_1^{*2} + Y_0^{*2} | 0, \widetilde{\boldsymbol{x}}\big) dP(\widetilde{\boldsymbol{x}}) = \boldsymbol{D}.
\end{aligned}$$

This finishes the proof. □

Proof of Lemma A.8 It suffices to prove the following two results:
(a) let $\boldsymbol{\theta}_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)/h_n$, then $\boldsymbol{\theta}_n = o_p(1)$;
(b) let $\{\boldsymbol{\beta}_n\} = \{(\beta_{n1}, \widetilde{\boldsymbol{\beta}}_n^T)^T\}$ be any sequence in $\mathbb{B}$ such that $(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)/h_n \to 0$ as $n \to \infty$, then $\boldsymbol{Q}_n(\boldsymbol{\beta}_n; h_n) \xrightarrow{p} \boldsymbol{Q}$.

To prove (a), we first note that $h_n\boldsymbol{\theta}_n \xrightarrow{p} 0$ by Theorem 2.1. Lemma A.3 then implies $\boldsymbol{R}_n(\boldsymbol{\theta}_n) \xrightarrow{p} 0$, and there exist some constants α_1 and α_2 such that: $||\boldsymbol{Q}\boldsymbol{\theta}_n|| \leqslant o_p(1) + \alpha_1 h_n||\boldsymbol{\theta}_n|| + \alpha_2 h_n||\boldsymbol{\theta}_n||^2$. Since $\boldsymbol{Q}$ is negative definite, we have $\inf_{\boldsymbol{\theta}} \frac{||\boldsymbol{Q}\boldsymbol{\theta}||}{||\boldsymbol{\theta}||} = |\omega_{min}| > 0$, where ω_{min} is the eigenvalue of $\boldsymbol{Q}$ with the smallest absolute value. It indicates that

$$0 < |\omega_{min}| < \frac{||\boldsymbol{Q}\boldsymbol{\theta}_n||}{||\boldsymbol{\theta}_n||} \leqslant o_p(||\boldsymbol{\theta}_n||^{-1}) + \alpha_1 h_n + \alpha_2 h_n||\boldsymbol{\theta}_n||.$$

Since $h_n \to 0$ and $h_n||\boldsymbol{\theta}_n|| \xrightarrow{p} 0$, if $||\boldsymbol{\theta}_n|| = o_p(1)$ does not hold, then the right hand side of the above inequality would degenerate to $o_p(1)$, which contradicts with the fact that it should be larger than $|\omega_{min}| > 0$. Consequently, we have $||\boldsymbol{\theta}_n|| = o_p(1)$.

To prove (b), let $q_{ni}(\boldsymbol{\beta}) = (2A_i - 1)K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\big)\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T Y_i$. It suffices to show that

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0|| \leqslant \epsilon h_n} \Big|\Big| n^{-1} \sum_{i=1}^n \big\{ h_n^{-2} q_{ni}(\boldsymbol{\beta}) - Q \big\} \Big|\Big| = o_p(1),$$

for arbitrary positive ϵ.

First, let $\widetilde{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}_0 + h_n\tilde{\boldsymbol{\theta}}_n$ with $\widetilde{\boldsymbol{\theta}}_n \to 0$ as $n \to \infty$. Now we have

$$\begin{aligned}
\mathrm{E}\big\{h_n^{-2}q_{ni}(\boldsymbol{\beta}_n)\big\} &= \mathrm{E}\Big\{2(2A-1)K''\big(\frac{x^T\boldsymbol{\beta}_n}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T}{h_n^2}Y\Big\} \\
&= \mathrm{E}\Big\{K''\big(\frac{x^T\boldsymbol{\beta}_0}{h_n} + \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}}\big)\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}'}{h_n^2}S(z,\widetilde{\boldsymbol{x}})\Big\} \\
&= \int K''\big(\frac{z}{h_n} + \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}}\big)\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T}{h_n^2}S(z,\widetilde{\boldsymbol{x}})f(z|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}}).
\end{aligned}$$

By Taylor expansion and (A3), there exists a $0 < \epsilon < 1$ such that $S(z,\widetilde{\boldsymbol{x}}) = zS^{(1)}(\epsilon z,\widetilde{\boldsymbol{x}})$. We have

$$\begin{aligned}
\mathrm{E}\big\{h_n^{-2}q_{ni}(\boldsymbol{\beta}_n)\big\} &= \int K''\big(\frac{z}{h_n} + \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}}\big)\frac{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T}{h_n^2}zS^{(1)}(\epsilon z,\widetilde{\boldsymbol{x}})f(z|\widetilde{\boldsymbol{x}})dzdP(\widetilde{\boldsymbol{x}}) \\
&= \int(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}})K''(\zeta)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TS^{(1)}\big(\epsilon h_n(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}}),\widetilde{\boldsymbol{x}}\big)f\big(h_n(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}})|\widetilde{\boldsymbol{x}}\big)d\zeta dP(\widetilde{\boldsymbol{x}}).
\end{aligned}$$

Then by (K1), (A2), (A4)-(A5), and the dominated convergence theorem,

$$\begin{aligned}
\lim_{n\to\infty}\mathrm{E}\big\{h_n^{-2}q_{ni}(\boldsymbol{\beta}_n)\big\} &= \lim_{n\to\infty}\int(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}})K''(\zeta)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TS^{(1)}\big(\epsilon h_n(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}}),\widetilde{\boldsymbol{x}}\big)f\big(h_n(\zeta - \widetilde{\boldsymbol{\theta}}_n^T\widetilde{\boldsymbol{x}})|\widetilde{\boldsymbol{x}}\big)d\zeta dP(\widetilde{\boldsymbol{x}}) \\
&= \lim_{n\to\infty}\int\zeta K''(\zeta)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TS^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})d\zeta dP(\widetilde{\boldsymbol{x}}) \\
&= a_2\int\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^TS^{(1)}(0,\widetilde{\boldsymbol{x}})f(0|\widetilde{\boldsymbol{x}})dP(\widetilde{\boldsymbol{x}}) = \boldsymbol{Q}.
\end{aligned}$$

The Donsker property of $\mathcal{Q}$ implies that for arbitrary $\epsilon > 0$,

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leqslant\epsilon h_n}\sup_{h\in(0,1]}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{q_{ni}(\boldsymbol{\beta}) - \mathrm{E}q_{ni}(\boldsymbol{\beta})\big\}\Big|\Big| = O_p(n^{-1/2}).$$

Then since $h_n = o(n^{-1/(2b+1)})$ and $(nh_n^4)^{-1} = o(1)$, we can derive that

$$\begin{aligned}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leqslant \epsilon h_n} \Big|\Big| n^{-1}\sum_{i=1}^{n}\big\{h_n^{-2}q_{ni}(\boldsymbol{\beta}) - Q\big\}\Big|\Big| \leqslant & \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leqslant \epsilon h_n} \Big|\Big| (nh_n^2)^{-1}\sum_{i=1}^{n}\big\{q_{ni}(\boldsymbol{\beta}) - \mathrm{E}q_{ni}(\boldsymbol{\beta})\big\}\Big|\Big| + o(1) \\ \leqslant & \; h_n^{-2}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||\leqslant \epsilon h_n}\sup_{h\in(0,1]} \Big|\Big| n^{-1}\sum_{i=1}^{n}\big\{q_{ni}(\boldsymbol{\beta}) - \mathrm{E}q_{ni}(\boldsymbol{\beta})\big\}\Big|\Big| \\ \leqslant & \; O_p(n^{-1/2}h_n^{-2}) = o_p(1).\end{aligned}$$

□

Proof of Lemma A.9 With the consistency result in Theorem 2.1, we have $P\big(\widehat{\beta}_{n1} = \beta_{01}\big) \to 1$ as $n \to \infty$. Hence for any $\boldsymbol{\beta} = (\beta_1, \widetilde{\boldsymbol{\beta}}^T)^T$ between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_n$, note that $\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0) = \mathrm{I}\big(z + \widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0) > 0\big)$. Then for $V(\boldsymbol{\beta}) = \mathrm{E}_X\{\mu(1,\boldsymbol{x})\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} > 0) + \mu(0,\boldsymbol{x})\mathrm{I}(\boldsymbol{x}^T\boldsymbol{\beta} \leqslant 0)\}$, we have

$$\begin{aligned}V(\boldsymbol{\beta}) &= \mathrm{E}_X\Big\{\mu(1,\boldsymbol{x})\mathrm{I}\big(z + \widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0) > 0\big) + \mu(0,\boldsymbol{x})\mathrm{I}\big(z + \widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0) \leqslant 0\big)\Big\} \\ &= \int\Big[\int_{-\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0)}^{\infty} \mathrm{E}_X\big\{\mu(1,\boldsymbol{x})|z,\widetilde{\boldsymbol{x}}\big\}f(z|\widetilde{\boldsymbol{x}})dz + \int_{-\infty}^{-\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_0)} \mathrm{E}_X\big\{\mu(0,\boldsymbol{x})|z,\widetilde{\boldsymbol{x}}\big\}f(z|\widetilde{\boldsymbol{x}})dz\Big]dP(\widetilde{\boldsymbol{x}}).\end{aligned}$$

Let $\widetilde{\boldsymbol{\delta}} = \widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0$, then $\widetilde{\boldsymbol{\delta}} \xrightarrow{p} \widetilde{\mathbf{0}}$ for $\boldsymbol{\beta}$ between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_n$, according to Theorem 2.1. Note that

$$\begin{aligned}V'(\boldsymbol{\beta}) &= \frac{\partial V(\boldsymbol{\beta})}{\partial\widetilde{\boldsymbol{\beta}}} = \int S\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}},\widetilde{\boldsymbol{x}}\big)f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T(\widetilde{\boldsymbol{\beta}}_0 + \widetilde{\boldsymbol{\delta}})dP(\widetilde{\boldsymbol{x}}), \\ V''(\boldsymbol{\beta}) &= \frac{\partial V'(\boldsymbol{\beta})}{\partial\widetilde{\boldsymbol{\beta}}} = \int S\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}},\widetilde{\boldsymbol{x}}\big)\Big\{f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big) - (\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0)f^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\Big\}\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dP(\widetilde{\boldsymbol{x}}) \\ &\quad - \int(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}})\Big\{S^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}},\widetilde{\boldsymbol{x}}\big)f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big) + S\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}},\widetilde{\boldsymbol{x}}\big)f^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\Big\}\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dP(\widetilde{\boldsymbol{x}}) \\ &\quad - \int(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0)S^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}},\widetilde{\boldsymbol{x}}\big)f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dP(\widetilde{\boldsymbol{x}}) \\ &= I_1 + I_2 + I_3,\end{aligned}$$

where the definition of I_i ($i = 1, 2, 3$) is clear from the context. By Taylor expansion, there exists some constant $0 < r_1 < 1$ such that

$$I_1 = \int -(\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}})S^{(1)}\big(-r_1\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}, \widetilde{\boldsymbol{x}}\big)\Big\{f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big) - (\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0)f^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\Big\}\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dP(\widetilde{\boldsymbol{x}}).$$

By (A2), (A4)-(A5), we know that the components of $\widetilde{\boldsymbol{x}}$, $S^{(i)}(z, \widetilde{\boldsymbol{x}})$ and $f^{(i)}(z|\widetilde{\boldsymbol{x}})$, $i = 0, 1$, are bounded for almost every $\widetilde{\boldsymbol{x}}$. Then for any $\widetilde{\boldsymbol{\delta}} \xrightarrow{p} \widetilde{\mathbf{0}}$, it is easy to conclude $I_1 \xrightarrow{p} 0$ and $I_2 \xrightarrow{p} 0$. To evaluate I_2, note that for some constant $0 < r_2 < 1$,

$$\begin{aligned} I_2 = \boldsymbol{I}_V + \int(\widetilde{\boldsymbol{\delta}}^T\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\beta}}_0)\Big\{ & S^{(2)}\big(-r_2\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}, \widetilde{\boldsymbol{x}}\big)f\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big) \\ + & S^{(1)}\big(-\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}, \widetilde{\boldsymbol{x}}\big)f^{(1)}\big(-r_2\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{\delta}}|\widetilde{\boldsymbol{x}}\big)\Big\}\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dP(\widetilde{\boldsymbol{x}}). \end{aligned}$$

With the boundedness of the components of $\widetilde{\boldsymbol{x}}$, $S^{(1)}(z, \widetilde{\boldsymbol{x}})$, $S^{(2)}(z, \widetilde{\boldsymbol{x}})$, $f(z|\widetilde{\boldsymbol{x}})$ and $f^{(1)}(z|\widetilde{\boldsymbol{x}})$ from (A2), (A4)-(A5), we also have $I_2 \xrightarrow{p} \boldsymbol{I}_V$ as $\widetilde{\boldsymbol{\delta}} \xrightarrow{p} \widetilde{\mathbf{0}}$, where $\boldsymbol{I}_V$ is negative definite by (A5). This finishes the proof. □

Recall from Section 2.3.2 that $r = \{r_1, \ldots, r_n\}$ denotes the collection of the random bootstrap weights and $w = \{W_1, \ldots, W_n\}$ denotes the random sample of observations, where $W_i = (\boldsymbol{x}_i, A_i, Y_i)$. Given a sequence of random variables R_n, $n = 1, \ldots, n$, we write $R_n = o_{p_r}(1)$ if for any $\epsilon > 0, \delta > 0$, we have $P_w\big(P_{r|w}(|R_n| > \epsilon) > \delta\big) \to 0$ as $n \to \infty$. In the bootstrap literature, R_n is said to converge to zero in probability, conditional on the data. Let $\mathrm{E}_{r|w}$ and $\mathrm{Var}_{r|w}$ denote the conditional expectation and the conditional variance according to the distribution of r given x. Furthermore, $o_{p_{rw}}(1)$ denotes a random sequence that converges to zero in probability with respect to the joint distribution of (r, w), and $o^*_{P_W}(1)$ denotes a random sequence that converges to zero in probability with respect to the distribution of r only. By Lemma 3 of Cheng and Huang (2010), if $R_n = o_{p_{rw}}(1)$, then $R_n = o_{p_r}(1)$. In particular, if R_n depends only on the data w but not on the random weights r and if $R_n = o_{p_w}(1)$, then it is easy to see $R_n = o_{p_{rw}}(1)$, and hence it is $o_{p_r}(1)$. In this

part of proof, we will include subscripts in the probability and expectation to clarify which probability distribution is used in the calculation.

Proof of Lemma A.10 It suffices to prove
(a) let $\boldsymbol{\theta}_n^* = (\widehat{\boldsymbol{\beta}}_n^* - \widehat{\boldsymbol{\beta}}_n)/h_n$, we have $\boldsymbol{\theta}_n^* = o_{p_r}(1)$;
(b) let $\{\boldsymbol{\beta}_n\} = \{(\beta_{n1}, \widetilde{\boldsymbol{\beta}}_n^T)^T\}$ be any sequence in $\mathbb{B}$ such that $(\boldsymbol{\beta}_n - \widehat{\boldsymbol{\beta}}_n)/h_n \to 0$ as $n \to \infty$, then $\boldsymbol{Q}_n^*(\boldsymbol{\beta}_n; h_n) = \boldsymbol{Q} + o_{p_r}(1)$.

To prove (a), for any $\boldsymbol{\theta} \in \mathbb{R}^{p-1}$, define $\boldsymbol{R}_n^*(\boldsymbol{\theta}) = \frac{2}{nh_n^2}\sum_{i=1}^n r_i(2A_i - 1)K'\big(\frac{z_i}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}_i\big)\widetilde{\boldsymbol{x}}_i Y_i$. We observe

$$\begin{aligned}\big|\big|\mathrm{E}_{r|w}\boldsymbol{R}_n^*(\boldsymbol{\theta}) - \boldsymbol{Q}\boldsymbol{\theta}\big|\big| &\leqslant \big|\big|\mathrm{E}_{r|w}\boldsymbol{R}_n^*(\boldsymbol{\theta}) - \mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta})\big|\big| + \big|\big|\mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta}) - \boldsymbol{Q}\boldsymbol{\theta}\big|\big| \\ &= \big|\big|R_n(\boldsymbol{\theta}) - \mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta})\big|\big| + \big|\big|\mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta}) - \boldsymbol{Q}\boldsymbol{\theta}\big|\big|.\end{aligned}$$

By Lemma A.3,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_n}\big|\big|\boldsymbol{R}_n(\boldsymbol{\theta}) - \mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta})\big|\big| &= o_p(1), \\ \big|\big|\mathrm{E}_w\boldsymbol{R}_n(\boldsymbol{\theta}) - \boldsymbol{Q}\boldsymbol{\theta}\big|\big| \leqslant o(1) + \alpha_1 h_n||\boldsymbol{\theta}|| &+ \alpha_2 h_n||\boldsymbol{\theta}||^2,\end{aligned}$$

uniformly over $\boldsymbol{\theta} \in \Theta_n$ for some finite α_1 and α_2. Hence

$$\big|\big|\mathrm{E}_{r|w}\boldsymbol{R}_n^*(\boldsymbol{\theta}) - \boldsymbol{Q}\boldsymbol{\theta}\big|\big| \leqslant o(1) + \alpha_1 h_n||\boldsymbol{\theta}|| + \alpha_2 h_n||\boldsymbol{\theta}||^2,$$

uniformly over $\boldsymbol{\theta} \in \Theta_n$. By Theorem 2.3, $h_n\boldsymbol{\theta}_n^* = o_{p_r}(1)$. So $\boldsymbol{R}_n^*(\boldsymbol{\theta}_n^*) = o_{p_r}(1)$. So we have

$$\big|\big|\boldsymbol{Q}\boldsymbol{\theta}_n^*\big|\big| \leqslant o(1) + \alpha_1 h_n||\boldsymbol{\theta}_n^*|| + \alpha_2 h_n||\boldsymbol{\theta}_n^*||^2.$$

Then similarly to the proof of Lemma B2, we can show that $\boldsymbol{\theta}_n^* = o_{p_r}(1)$.

To prove (b), let $q^*_{ni}(\boldsymbol{\beta}) = (2A_i - 1)rK''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\big)\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T Y_i$. It suffices to show that

$$\sup_{||\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_n||\leqslant\epsilon h_n}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{h_n^{-2}q^*_{ni}(\boldsymbol{\beta})-Q\big\}\Big|\Big| = o_{p_{rw}}(1),$$

for arbitrary positive ϵ.

Let $\widetilde{\boldsymbol{\beta}}^r_n = \widetilde{\boldsymbol{\beta}}_n + h_n\widetilde{\boldsymbol{\theta}}^*_n$ with $\widetilde{\boldsymbol{\theta}}^*_n \to 0$. Consequently, $\lim_{n\to\infty}\mathrm{E}_w\mathrm{E}_{r|w}\{h_n^{-2}q^*_{ni}(\boldsymbol{\beta}^r_n)\} = Q$.

The Donsker property of $\mathcal{Q}^*$ implies that for arbitrary $\epsilon > 0$,

$$\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_n||\leqslant\epsilon h_n}\sup_{h\in(0,1]}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{q^*_{ni}(\boldsymbol{\beta})-\mathrm{E}_w\mathrm{E}_{r|w}q^*_{ni}(\boldsymbol{\beta})\big\}\Big|\Big| = O_{p_{rw}}(n^{-1/2}).$$

Then since $h_n = o(n^{-1/(2b+1)})$ and $(nh_n^4)^{-1} = o(1)$, we can derive that

$$\begin{aligned}
\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_n||\leqslant\epsilon h_n} n^{-1}\Big|\Big|\sum_{i=1}^n\big\{h_n^{-2}q^*_{ni}(\boldsymbol{\beta})-Q\big\}\Big|\Big| &\leqslant \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_n||\leqslant\epsilon h_n}(nh_n^2)^{-1}\Big|\Big|\sum_{i=1}^n\big\{q^*_{ni}(\boldsymbol{\beta})-\mathrm{E}_w\mathrm{E}_{r|w}q^*_{ni}(\boldsymbol{\beta})\big\}\Big|\Big| + o(1)\\
&\leqslant h_n^{-2}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_n||\leqslant\epsilon h_n}\sup_{h\in(0,1]}\Big|\Big|n^{-1}\sum_{i=1}^n\big\{q^*_{ni}(\boldsymbol{\beta})-\mathrm{E}_w\mathrm{E}_{r|w}q^*_{ni}(\boldsymbol{\beta})\big\}\Big|\Big|\\
&\leqslant O_{p_{rw}}(n^{-1/2}h_n^{-2}) = o_{p_{rw}}(1).
\end{aligned}$$

□

A.7 Moving Parameter Asymptotics

To better understand the behavior of the proposed inference procedure, we study the properties of the smoothed estimator and its bootstrapped version under a moving parameter or local asymptotic framework, as motivated by Laber et al. (2014).

Consider the following semiparametric model

$$Y = \mu(\boldsymbol{x}) + \boldsymbol{x}^T(\boldsymbol{\beta}_0 + b_n\boldsymbol{s})\boldsymbol{x}A + \epsilon, \tag{A.9}$$

where $\mu(\boldsymbol{x})$ is an unspecified function, ϵ is a sub-Gaussian random error term with mean zero and variance σ^2. The local model (A.9) perturbs $\boldsymbol{\beta}_0 = (\beta_{01}, \widetilde{\boldsymbol{\beta}}_0)$ (with $|\beta_{01}| = 1$) by a small quantity $b_n\boldsymbol{s}$, with b_n being a sequence of real numbers that converges to zero as $n \to \infty$ and $\boldsymbol{s} = (s_1, \widetilde{\boldsymbol{s}})$ is a fixed p-dimensional vector. We write $\boldsymbol{s} = (s_1, \widetilde{\boldsymbol{s}})$ and assume $s_1 = 0$ to avoid complications that are not relevant to the main results. When $b_n = 0$, the optimal treatment regime is given by $I(\boldsymbol{x}^T\boldsymbol{\beta}_0 > 0)$.

Consider a random sample $\{(\boldsymbol{x}_i, A_i, Y_i), i = 1, ..., n\}$ from (A.9). We estimate $\boldsymbol{\beta}_0$ by the smooth robust estimator introduced in Section 2.2.3, that is, $\widehat{\boldsymbol{\beta}}_n = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n (2A_i - 1)K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)Y_i$. Correspondingly, the confidence interval is constructed using the formula in (2.8) based on the bootstrapped estimator $\widehat{\boldsymbol{\beta}}_n^* = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} n^{-1}\sum_{i=1}^n r_i(2A_i - 1)K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)Y_i$. That is, we study the behavior of the procedures proposed earlier which are constructed in a model-free fashion when the underlying data are generated by (A.9). To study the local asymptotics, define

$$\boldsymbol{D}_0 = 2a_1\mathrm{E}\big[\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}})\big\{\mathrm{E}(\mu^2(\boldsymbol{x})|z = 0,\ \widetilde{\boldsymbol{x}}) + \sigma^2\big\}\big] \quad \text{and} \quad \boldsymbol{Q}_0 = a_2\mathrm{E}\big\{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T f(0|\widetilde{\boldsymbol{x}})\big\},$$

where a_i ($i = 1, 2$) is defined in Section 2.3.1. As before, write $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}, \widetilde{\boldsymbol{\beta}}_n^T)^T \in \mathbb{R}^p$ and $\widehat{\boldsymbol{\beta}}_n^* = (\widehat{\beta}_{n1}^*, \widetilde{\boldsymbol{\beta}}_n^{*T})^T$.

The following two theorems show that asymptotic normality holds for $\widehat{\boldsymbol{\beta}}_n$ and $\widetilde{\boldsymbol{\beta}}_n^{*T}$ for b_n chosen at appropriate rate. If the sequence b_n goes to zero faster that $(nh_n)^{-1/2}$, the smoothed estimator is asymptotically unbiased and the bootstrap confidence interval for $\boldsymbol{\beta}_0$ is asymptotically accurate. The proofs of these results can be found in Appendix A.8.

Theorem A.1

Assume $K(\cdot)$ satisfies (K1) - (K3) for some $b \geqslant 2$, $h_n = o(n^{-1/(2b+1)})$ and $n^{-1}h_n^{-4} = o(1)$. If $b_n = (nh_n)^{-1/2}$, then under (A1), (A2), (A4),

$$\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_0) \to N\big(-a_2^{-1}\widetilde{\boldsymbol{s}}, \boldsymbol{Q}_0^{-1}\boldsymbol{D}_0\boldsymbol{Q}_0^{-1}\big)$$

in distribution as $n \to \infty$. □

Theorem A.2

Assume $K(\cdot)$ satisfies (K1) - (K3), for some $b \geqslant 2$, $h_n = o(n^{-1/(2b+1)})$, and $\log(n) = o(nh_n^4)$. If $b_n = (nh_n)^{-1/2}$, then under (A1), (A2), (A4), (A6),

$$\sqrt{nh_n}(\widetilde{\boldsymbol{\beta}}_n^* - \widetilde{\boldsymbol{\beta}}_n) = N\big(-a_2^{-1}\widetilde{\boldsymbol{s}}, \boldsymbol{Q}_0^{-1}\boldsymbol{D}_0\boldsymbol{Q}_0^{-1}\big) + o_{p_r}(1). \qquad \square$$

A.8 Proof of Results in Appendix A.7

Let Y be the response generated from the local model (A.9). We can write $Y = \breve{Y} + b_n\widetilde{Y}$, where $\breve{Y} = \mu(\boldsymbol{x}) + A\boldsymbol{x}^T\boldsymbol{\beta}_0 + \epsilon$ and $\widetilde{Y} = A\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{s}}$. Note that $\breve{Y}$ satisfied all the assumptions about the outcome variable in (A1), (A3) and (A5). It follows that all the preceding lemmas and theorems still hold if regarding $\breve{Y}$ as the observed response Y. In addition, since $(2A-1)\widetilde{Y}$ is square integrable and does not depend on $(\boldsymbol{\beta}, h)$, it implies that all classes listed in Lemma A.1 are still VC classes with $\widetilde{Y}$ as their responses. In the following proof, we use "$\breve{}$" to denote corresponding notation when we replace Y with $\breve{Y}$. For example, we define

$$\breve{\widetilde{M}}_n(\boldsymbol{\beta}, h_n) = 2n^{-1}\sum_{i=1}^{n}(2A_i - 1)\mathrm{I}(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0)\breve{Y}.$$

First we will prove the consistency of the smoothed estimator given the observed data $\{(\boldsymbol{x}_i, A_i, Y_i), i = 1, ..., n\}$ from (A.9).

Lemma A.11

Under (A1), (A2) and assume $K(\cdot)$ satisfies (K1), if $b_n = o(1)$, then $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + o_p(1)$. □

Proof of Lemma A.11 We observe that $\widehat{\boldsymbol{\beta}}_n$ maximizes $\widetilde{M}_n(\boldsymbol{\beta}, h_n)$ over $\boldsymbol{\beta} \in \mathbb{B}$, and $\boldsymbol{\beta}_0$ maximizes $\breve{M}(\boldsymbol{\beta})$. Note that

$$\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \breve{M}(\boldsymbol{\beta})\big| \leqslant \sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|2b_n n^{-1} \sum_{i=1}^{n} (2A_i - 1) K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{Y}_i\big| + \sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\widetilde{\breve{M}}_n(\boldsymbol{\beta}, h_n) - \breve{M}(\boldsymbol{\beta})\big|.$$

Lemma A.2 implies that $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\widetilde{\breve{M}}_n(\boldsymbol{\beta}, h_n) - \breve{M}(\boldsymbol{\beta})\big| = o_p(1)$. In addition, it is obvious that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\frac{b_n}{n} \sum_{i=1}^{n} (2A_i - 1) K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{Y}_i\,\big| \;\leqslant\; & \sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big|\frac{b_n}{n} \sum_{i=1}^{n} \Big[(2A_i - 1) K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{Y}_i - \mathrm{E}\big\{K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}}\big\}\Big]\Big| \\ & + \sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|b_n \mathrm{E}\big\{K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}}\big\}\big|.\end{aligned}$$

The Donsker property of $\mathcal{F}$ ensures the first term converges to 0 in probability if $b_n = o(\sqrt{n})$. By the boundedness of $K(\cdot)$ and $\boldsymbol{x}$, the second term also goes to 0 as $b_n = o(1)$. So $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \breve{M}(\boldsymbol{\beta})\big| = o_p(1)$ can be concluded.

The construction of $\breve{Y}$ implies that for every $\tau > 0$,

$$\begin{aligned}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\tau} \breve{M}(\boldsymbol{\beta}) - \breve{M}(\boldsymbol{\beta}_0) &= \sup_{\boldsymbol{\beta}\in\mathbb{B}} 2\mathrm{E}[(2A_i - 1)\breve{Y}_i\{I(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) - I(\boldsymbol{x}_i^T\boldsymbol{\beta}_0 > 0)\}] \\ &= \sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}_0||>\tau} \mathrm{E}[\boldsymbol{x}_i^T\boldsymbol{\beta}_0\{(\boldsymbol{x}_i^T\boldsymbol{\beta} > 0) - I(\boldsymbol{x}_i^T\boldsymbol{\beta}_0 > 0)\}] < 0.\end{aligned}$$

Hence, $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + o_p(1)$ is derived from Theorem 5.7 in van der Vaart (2000). □

Proof of Theorem A.1 The proof of Theorem 2.2 implies that it suffices to verify:

(a) $(nh_n)^{1/2}\boldsymbol{T}_n(\boldsymbol{\beta}_0; h_n) \xrightarrow{d} N(a_2^{-1}\boldsymbol{Q}_0\widetilde{\boldsymbol{s}}, \boldsymbol{D}_0)$;

(b) $\boldsymbol{Q}_n(\boldsymbol{\beta}_n^r; h_n) = \boldsymbol{Q}_0 + o_p(1)$ for any $\boldsymbol{\beta}_n^r$ is between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$.

To prove (a), note that Lemma A.7 indicates that

$$\mathrm{E}(nh_n)^{1/2}\big\{\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\} \to a_2^{-1}\boldsymbol{Q}_0\widetilde{\boldsymbol{s}}, \quad \text{and} \quad \mathrm{Var}(nh_n)^{1/2}\big\{\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\} \to \boldsymbol{D}_0.$$

It is sufficient to prove that $(nh_n)^{1/2}\boldsymbol{\gamma}^T\big\{\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n) - \mathrm{E}\boldsymbol{T}_n(\boldsymbol{\beta}_0;h_n)\big\}$ is asymptotically $N(\widetilde{\mathbf{0}}, \boldsymbol{\gamma}^T\boldsymbol{D}_0\boldsymbol{\gamma})$ for any fixed vector $\boldsymbol{\gamma} \in \mathbb{R}^{p-1}$ such that $||\boldsymbol{\gamma}|| = 1$. Define

$$\begin{aligned} q_i &= 2(2A_i-1)(nh_n)^{1/2}K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i}{h_n}Y_i \\ &= \breve{q}_i + 2(2A_i-1)K'\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}_0}{h_n}\big)\frac{\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i}{h_n}\widetilde{Y}_i \\ &= \breve{q}_i + \widetilde{q}_i \end{aligned}$$

for $\breve{q}_i$ defined as in the proof of Theorem 2.2. With Lyapunov central limit theorem, we will verify

$$\lim_{n\to\infty}(s_n)^{-4}\sum_{i=1}^{n}\mathrm{E}\big\{(q_i - \mathrm{E}q_i)^4\big\} = 0, \tag{A.10}$$

where $\lim_{n\to\infty}(n^{-1}s_n)^2 = \lim_{n\to\infty}\sum_{i=1}^{n}\mathrm{Var}(n^{-1}q_i) = \boldsymbol{\gamma}^T\boldsymbol{D}_0\boldsymbol{\gamma}$. The fact that $\lim_{n\to\infty} n^{-3}\mathrm{E}\big\{(\breve{q}_i - \mathrm{E}\breve{q}_i)^4\big\} = 0$ implies that the left-side of (A.10) is bounded from above (up to a positive constant) by

$$\begin{aligned} \lim_{n\to\infty} n^{-3}\mathrm{E}(\widetilde{q}_i^4) + \lim_{n\to\infty} n^{-3}(\mathrm{E}\widetilde{q}_i)^4 &= \lim_{n\to\infty}(n^3h_n^4)^{-1}8\mathrm{E}\big\{K'(\frac{x_i^T\boldsymbol{\beta}_0}{h_n})^4(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}})^4\big\} \\ &\quad + \lim_{n\to\infty}(n^3h_n^4)^{-1}\big[\mathrm{E}\big\{K'(\frac{x_i^T\boldsymbol{\beta}_0}{h_n})(\boldsymbol{\gamma}^T\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}})\big\}\big]^4. \end{aligned}$$

With the boundedness of $K'(\cdot)$ and $\widetilde{\boldsymbol{x}}$, $(n^3h_n^4)^{-1} = o(1)$ implies the Lyapunov condition is

satisfied, and (a) follows. To prove (b), note that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{B}}||\boldsymbol{Q}_n(\boldsymbol{\beta};h_n)-\boldsymbol{Q}_n(\boldsymbol{\beta};h_n)|| &= \sup_{\boldsymbol{\beta}\in\mathbb{B}}\Big|\Big|\frac{2b_n}{n}\sum_{i=1}^{n}(2A_i-1)K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n^2}\widetilde{Y}_i\Big|\Big| \\ &\leqslant O_p\big((n^2h_n^5)^{-1/2}\big)+O(b_nh_n^{-1})\sup_{\boldsymbol{\beta}\in\mathbb{B}}\Big|\Big|\mathrm{E}\big\{K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n}\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{s}}\big\}\Big|\Big| \\ &= O_p\big((n^2h_n^5)^{-1/2}\big)+O\big((nh_n^3)^{-1/2}\big)=o_p(1),\end{aligned}$$

since $\mathcal{Q}$ is a VC class. Then it suffices to show that $\boldsymbol{\theta}_n=(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)/h_n=o_p(1)$. Consider $\boldsymbol{R}_n(\boldsymbol{\theta})$ defined as in Lemma A.3. The Donsker properties of $\mathcal{H}$ imply that

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_n}||\boldsymbol{R}_n(\boldsymbol{\theta})-\breve{\boldsymbol{R}}_n(\boldsymbol{\theta})|| &\leqslant O_p\big((n^2h_n^5)^{-1/2}\big)+O(b_nh_n^{-1})\sup_{\boldsymbol{\theta}\in\Theta_n}\Big|\Big|\mathrm{E}\big\{K'\big(\frac{z_i}{h_n}+\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}_i\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}}}{h_n}\big\}\Big|\Big| \\ &= O_p\big((n^2h_n^5)^{-1/2}\big)+O\big((nh_n^3)^{-1/2}\big)=o_p(1),\end{aligned}$$

where Θ_n is defined in Lemma A3. Combined with Lemma A3, it implies that $\sup_{\boldsymbol{\theta}\in\Theta_n}||\boldsymbol{R}_n(\boldsymbol{\theta})-\boldsymbol{Q}_0\boldsymbol{\theta}||\leqslant o(1)+\alpha_1h_n||\boldsymbol{\theta}||+\alpha_2h_n||\boldsymbol{\theta}||^2$. By the definition of $\boldsymbol{\theta}_n$, we know that $h_n\boldsymbol{\theta}_n\xrightarrow{p}0$, and $\boldsymbol{R}_n(\boldsymbol{\theta}_n)\xrightarrow{p}0$. Then from the proof of Lemma A.8, (b) can be concluded. □

For the asymptotic distribution for bootstrap estimators with the moving parameter framework, we first prove its consistency.

Lemma A.12

Under (A1), (A2), (A6) and assume $K(\cdot)$ satisfies (K1), if $b_n=o(1)$, then $\widehat{\boldsymbol{\beta}}_n^*=\widehat{\boldsymbol{\beta}}_n+o_{p_r}(1)$. □

Proof of Lemma A.12 By definition, $\widehat{\boldsymbol{\beta}}_n^*$ maximizes $\widetilde{M}_n^*(\boldsymbol{\beta}, h_n)$ over $\boldsymbol{\beta} \in \mathbb{B}$. First, given $\mathcal{F}^{*new}$ is a VC class, the Donsker property and Lemma A.2 jointly indicate that

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\widetilde{M}_n^*(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\boldsymbol{\beta}, h_n)\right| \;\leqslant\; & \sup_{\boldsymbol{\beta}\in\mathbb{B}}\Big|\frac{2b_n}{n}\sum_{i=1}^{n}(r_i-1)(2A_i-1)K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\widetilde{Y}_i\Big| \\ & + \sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\widetilde{\widetilde{M}}_n^*(\boldsymbol{\beta}, h_n) - \widetilde{\widetilde{M}}_n(\boldsymbol{\beta}, h_n)\right| = o_{p_{rw}}(1).\end{aligned}$$

By Lemma 3 of Cheng & Huang (2010), $\sup_{\boldsymbol{\beta}\in\mathbb{B}}\left|\widetilde{M}_n^*(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\boldsymbol{\beta}, h_n)\right| = o_{p_r}(1)$. By Theorem 5.7 in van der Vaart (2000), to prove the theorem, it is sufficient to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P_w\Big(\sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon}\big\{\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\widehat{\boldsymbol{\beta}}_n, h_n)\big\} < 0\Big) = 1. \tag{A.11}$$

Note that Lemma A.2 and the consistency of $\widehat{\boldsymbol{\beta}}_n$ implies that

$$\begin{aligned}& \sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon}\big\{\widetilde{M}_n(\boldsymbol{\beta}, h_n) - \widetilde{M}_n(\widehat{\boldsymbol{\beta}}_n, h_n)\big\} \\ \leqslant\; & \sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon}\big\{\widetilde{\widetilde{M}}_n(\boldsymbol{\beta}, h_n) - \widetilde{\widetilde{M}}_n(\boldsymbol{\beta}_0, h_n)\big\} + \big\{\widetilde{\widetilde{M}}_n(\boldsymbol{\beta}_0, h_n) - \widetilde{\widetilde{M}}_n(\widehat{\boldsymbol{\beta}}_n, h_n)\big\} \\ & + \sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon}\Big|\frac{2b_n}{n}\sum_{i=1}^{n}(2A_i-1)\Big\{K\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big) - K\big(\frac{\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_n}{h_n}\big)\Big\}\widetilde{Y}_i\Big| \\ =\; & \sup_{||\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_n||>\epsilon}\big\{\widetilde{M}(\boldsymbol{\beta}) - \widetilde{M}(\boldsymbol{\beta}_0)\big\} + o_p(1).\end{aligned}$$

Furthermore, the consistency of $\widehat{\boldsymbol{\beta}}_n$ implies that for all sufficiently large n, any $\boldsymbol{\beta}$ that satisfies $||\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_n|| > \tau$ would also satisfy $||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \geqslant \tau/2$. Hence, Lemma A.11 implies (A.11) holds. □

Proof of Theorem A.2 The proofs of Theorem 2.4 indicate that it is sufficient to verify:
(a) $\boldsymbol{Q}_n^*(\boldsymbol{\beta}_n^{*r}; h_n) = \boldsymbol{Q}_0 + o_{p_r}(1)$ for any $\boldsymbol{\beta}_n^{*r}$ is between $\widehat{\boldsymbol{\beta}}_n^*$ and $\widehat{\boldsymbol{\beta}}_n$;
(b) $(nh_n)^{1/2}\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n; h_n) = N(a_2^{-1}\boldsymbol{Q}_0\widetilde{\boldsymbol{s}}, \boldsymbol{D}_0) + o_{p_r}(1)$.

To prove (a), the fact that $\mathcal{Q}^*$ is a VC class implies that

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathbb{B}} \big|\big|\boldsymbol{Q}_n^*(\boldsymbol{\beta}; h_n) - \breve{\boldsymbol{Q}}_n^*(\boldsymbol{\beta}; h_n)\big|\big| &= \sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big|\Big| \frac{2b_n}{n}\sum_{i=1}^n r_i(2A_i - 1)K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n^2}\widetilde{Y}_i\Big|\Big| \\
&\leqslant O_p\big((n^2h_n^5)^{-1/2}\big) + O(b_nh_n^{-1}) \sup_{\boldsymbol{\beta}\in\mathbb{B}} \Big|\Big| \mathrm{E}_w\big\{K''\big(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{h_n}\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T}{h_n}\widetilde{\boldsymbol{x}}^T\widetilde{\boldsymbol{s}}\big\}\Big|\Big| \\
&= O_p\big((n^2h_n^5)^{-1/2}\big) + O\big((nh_n^3)^{-1/2}\big) = o_{p_r}(1).
\end{aligned}$$

It suffices to show that $\boldsymbol{\theta}_n^* = (\widehat{\boldsymbol{\beta}}_n^* - \boldsymbol{\beta}_0)/h_n = o_{p_r}(1)$. For $\boldsymbol{R}_n^*(\boldsymbol{\theta})$ defined in the proof of Lemma A.10, the fact that $\mathcal{H}^*$ is a VC class indicates that

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta_n} ||\boldsymbol{R}_n^*(\boldsymbol{\theta}) - \breve{\boldsymbol{R}}_n^*(\boldsymbol{\theta})|| &\leqslant O_p\big((n^2h_n^5)^{-1/2}\big) + O(b_nh_n^{-1}) \sup_{\boldsymbol{\theta}\in\Theta_n} \Big|\Big| \mathrm{E}_w\big\{K'\big(\frac{z_i}{h_n} + \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}_i\big)\frac{\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i^T\widetilde{\boldsymbol{s}}}{h_n}\big\}\Big|\Big| \\
&= O_p\big((n^2h_n^5)^{-1/2}\big) + O\big((nh_n^3)^{-1/2}\big) = o_{p_r}(1),
\end{aligned}$$

where Θ_n is defined in Lemma A3. Combined with Lemma B4, it implies that $\sup_{\boldsymbol{\theta}\in\Theta_n} ||\boldsymbol{R}_n^*(\boldsymbol{\theta}) - \boldsymbol{Q}_0\boldsymbol{\theta}|| \leqslant o(1) + \alpha_1 h_n||\boldsymbol{\theta}|| + \alpha_2 h_n||\boldsymbol{\theta}||^2$. By the definition of $\boldsymbol{\theta}_n^*$, we know that $h_n\boldsymbol{\theta}_n^* = o_{p_r}(1)$, and $\boldsymbol{R}_n^*(\boldsymbol{\theta}_n^*) = o_{p_r}(1)$. Then from the proof of Lemma B4, (a) can be concluded. To prove (b), the proof of Lemma A5 and the VC class $\mathcal{H}^*$ implies that $(nh_n)^{1/2}\{\boldsymbol{T}_n^*(\widehat{\boldsymbol{\beta}}_n; h_n) - \boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\} = o_{p_r}(1)$. Then observe that

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E}_w\mathrm{E}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} &= a_2^{-1}\boldsymbol{Q}_0\widetilde{\boldsymbol{s}}, \\
\lim_{n\to\infty} \mathrm{E}_w\Big[\,\mathrm{Var}_{r|w}\big\{(nh_n)^{1/2}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\}\Big] &= \boldsymbol{D}_0.
\end{aligned}$$

It is sufficient to prove that $(nh_n)^{1/2}\boldsymbol{\gamma}^T\big\{\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n) - \mathrm{E}\boldsymbol{T}_n^*(\boldsymbol{\beta}_0; h_n)\big\} = N(\widetilde{\boldsymbol{0}}, \boldsymbol{\gamma}^T\boldsymbol{D}_0\boldsymbol{\gamma}) +$

$o_{p_r}(1)$ for any fixed vector $\boldsymbol{\gamma} \in \mathbb{R}^{p-1}$ such that $||\boldsymbol{\gamma}|| = 1$. Define

$$\begin{aligned} q_i^* &= 2r_i(2A_i - 1)(nh_n)^{1/2} K'\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}_0}{h_n}\big) \frac{\boldsymbol{\gamma}^T \widetilde{\boldsymbol{x}}_i}{h_n} Y_i \\ &= \breve{q}_i^* + 2r_i(2A_i - 1) K'\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}_0}{h_n}\big) \frac{\boldsymbol{\gamma}^T \widetilde{\boldsymbol{x}}_i}{h_n} \widetilde{Y}_i = \breve{q}_i^* + \widetilde{q}_i^*, \end{aligned}$$

for $\breve{q}_i^*$ defined as in the proof of Theorem 4. To check the Lyapunov condition, it suffices to prove

$$\lim_{n\to\infty} (s_n^*)^{-4} \sum_{i=1}^n \mathrm{E}\big\{(q_i^* - \mathrm{E}q_i^*)^4\big\} = 0,$$

where $(s_n^*)^2 = \sum_{i=1}^n \mathrm{Var}_{r|w}(q_i^*)$. Similarly as the proof of Theorem A.1, the Lyapunov condition holds if

$$(s_n^*)^{-4} \sum_{i=1}^n \mathrm{E}_{r|w}(\widetilde{q}_i^{*4}) \xrightarrow{a.s.} 0, \quad \text{and} \quad (s_n^*)^{-4} \sum_{i=1}^n (\mathrm{E}_{r|w} \widetilde{q}_i^*)^4 \xrightarrow{a.s.} 0.$$

Since r is sub-Gaussian, then $\mathrm{E}|r|^k$ is finite for any positive integer k. Hence with bounded $K(\cdot)$ and $\widetilde{\boldsymbol{x}}$ and fixed $\boldsymbol{s}$, the strong law of large numbers and the continuous mapping theorem imply that

$$\begin{aligned} (s_n^*)^{-4} \sum_{i=1}^n \mathrm{E}_{r|w}(\widetilde{q}_i^{*4}) &\xrightarrow{a.s.} \{\lim_{n\to\infty} \mathrm{E}_w (s_n^*)^2\}^{-2} \lim_{n\to\infty} \mathrm{E}_w \sum_{i=1}^n \mathrm{E}_{r|w}(\widetilde{q}_i^{*4}) = 0, \\ (s_n^*)^{-4} \sum_{i=1}^n (\mathrm{E}_{r|w} \widetilde{q}_i^*)^4 &\xrightarrow{a.s.} \{\lim_{n\to\infty} \mathrm{E}_w (s_n^*)^2\}^{-2} \lim_{n\to\infty} \mathrm{E}_w \sum_{i=1}^n (\mathrm{E}_{r|w} \widetilde{q}_i^*)^4 = 0. \end{aligned}$$

This verifies the Lyapunov condition and (b) follows. □

A.9 Pseudo Codes for the Proximal Algorithm

Algorithm 2 Proximal ($\boldsymbol{\beta}^{(0)}$, α_0, γ)

Set $t = 0$.
Set diff = 0.
while diff $\geqslant 0$ **do**
    $t \leftarrow t + 1$.
    $\alpha_t \leftarrow \gamma\alpha_{t-1}$.
    $\delta_t \leftarrow (n\alpha_t)^{-1} \sum_{i=1}^{n} (2A_i - 1) K'\big(\frac{\boldsymbol{x}_i^T \beta^{(t-1)}}{h_n}\big) \frac{\boldsymbol{x}_i}{h_n} Y_i$.
    $\boldsymbol{\beta}^{(t)} \leftarrow \beta^{(t-1)} + \delta_t$.
    diff $\leftarrow \frac{2}{n} \sum_{i=1}^{n} (2A_i - 1) Y_i \big\{ K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)}}{h_n}\big) - K\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)}}{h_n}\big) - h_n^{-1} \boldsymbol{x}_i^T \delta_t K'\big(\frac{\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t-1)}}{h_n}\big) \big\} + \alpha_t ||\delta_t||^2$.
end while
Output $\boldsymbol{\beta}^{(t)}$.

A.10 Additional numerical results

Example A1 (binary response). The binary response Y is generated to satisfy $\mathrm{E}Y = (1+e^{-\boldsymbol{x}^T \boldsymbol{\beta}^{opt}})^{-1}$, where $\boldsymbol{X}$, $\boldsymbol{\beta}^{opt}$, A are the same as in Settings 1 & 2 in Chapter 2. Table A.1 and Table A.2 summarize the simulations results. We observe satisfactory performance as in the continuous response cases.

Table A.1: Monte Carlo estimates of the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value with binary outcomes.

n	β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	Match Ratio	$V_n(\widehat{\boldsymbol{\beta}}_n)$
			Setting 1			
300	0.03 (0.33)	0 (0)	0.03 (0.30)	0.03 (0.35)	99.60%	0.00 (0.17)
500	0.01 (0.21)	0 (0)	0.02 (0.20)	0.02 (0.22)	99.75%	0.00 (0.13)
1000	0.02 (0.13)	0 (0)	0.01 (0.13)	0.01 (0.15)	99.73%	-0.01 (0.09)
			Setting 2			
300	-0.04 (0.25)	0 (0)	0.02 (0.24)	0.00 (0.17)	99.40%	-0.01 (0.15)
500	-0.03 (0.20)	0 (0)	0.02 (0.19)	0.00 (0.13)	99.58%	-0.01 (0.12)
1000	0.00 (0.13)	0 (0)	0.01 (0.13)	0.00 (0.09)	99.90%	0.00 (0.08)

Table A.2: Empirical coverage probabilities and average interval lengths of the 95% bootstrap confidence intervals for $\boldsymbol{\beta}^{opt}$ with binary outcomes.

n		β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	$V(\boldsymbol{\beta}^{opt})$
Setting 1						
300	Coverage Rate	91.1%	100%	91.2%	90.5%	94.5%
	Average Length	1.35	0	1.24	1.38	0.69
500	Coverage Rate	94.6%	100%	93.1%	93.3%	94.3%
	Average Length	0.85	0	0.82	0.88	0.54
1000	Coverage Rate	94.6%	100%	95.1%	93.8%	96.0%
	Average Length	0.56	0	0.55	0.58	0.38
Setting 2						
300	Coverage Rate	93.5%	100%	93.8%	97.6%	94.4%
	Average Length	1.11	0	1.03	0.72	0.62
500	Coverage Rate	92.6%	100%	94.3%	96.0%	93.7%
	Average Length	0.78	0	0.74	0.53	0.48
1000	Coverage Rate	94.4%	100%	94.6%	95.7%	93.6%
	Average Length	0.52	0	0.50	0.36	0.34

Example A2 (different choices of kernel function). We consider the same data generative model as in Settings 1 & 2 in Chapter 2, and evaluate two different choices of kernels $K(\cdot)$. The first choice uses $K_1(\cdot) = \Phi(\cdot)$, the cumulative distribution function of standard normal distribution. The second choice is $K_2(v) = \big[0.5 + \frac{105}{64}\{\frac{v}{5} - \frac{5}{3}(\frac{v}{5})^3 + \frac{7}{5}(\frac{v}{5})^5 - \frac{3}{7}(\frac{v}{5})^7\}\big] I(-5 \leqslant v \leqslant 5) + I(v > 5)$. Its bandwidth is selected by $h_n = 0.9n^{-1/9} \min\{\text{ std}(\boldsymbol{x}_i^T\boldsymbol{\beta}),\ \text{IQR}(\boldsymbol{x}_i^T\boldsymbol{\beta})/1.34\}$. Both choices satisfy the regularity conditions in Appendix A.2. The bandwidth was chosen the same way as described in Section 2.4. The simulation results are summarized in Table A.3. We observe that the performance is not sensitive to different choices of kernel functions.

Example A3 (observational data). The data generative model is the same as that in in Settings 1 & 2 Chapter 2, except that A is generated according to $P(A = 1|\boldsymbol{x}) = \{1 + \exp(-\boldsymbol{x}^T\boldsymbol{\eta})\}^{-1}$, where $\boldsymbol{\eta} = (0.2, 0.5, 0.5, 0.5)^T$. Table A.4 summarizes the performance of the propensity score inverse weighted estimator given in (2.13) of Chapter 2. We observed

Table A.3: Monte Carlo estimates of the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value with with different choices of $K(\cdot)$.

n	Kernel	β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	Match Ratio	$V_n(\widehat{\boldsymbol{\beta}}_n)$
				Setting 1			
300	K_1	-0.05 (0.30)	0 (0)	0.01 (0.27)	0.04 (0.31)	99.35%	-0.02 (0.17)
	K_2	-0.05 (0.27)	0 (0)	0.04 (0.26)	0.05 (0.28)	99.40%	-0.01 (0.16)
500	K_1	-0.01 (0.19)	0 (0)	0.01 (0.20)	0.02 (0.22)	99.73%	0.00 (0.13)
	K_2	-0.03 (0.21)	0 (0)	0.02 (0.20)	0.03 (0.22)	99.60%	-0.01 (0.13)
1000	K_1	-0.01 (0.14)	0 (0)	0.00 (0.13)	0.01 (0.15)	99.88%	-0.01 (0.09)
	K_2	-0.01 (0.14)	0 (0)	0.02 (0.14)	0.02 (0.14)	99.77%	0.00 (0.09)
				Setting 2			
300	K_1	0.04 (0.26)	0 (0)	0.02 (0.24)	0.02 (0.18)	99.35%	-0.01 (0.15)
	K_2	-0.03 (0.24)	0 (0)	0.02 (0.24)	0.00 (0.17)	99.47%	-0.01 (0.14)
500	K_1	0.02 (0.19)	0 (0)	0.02 (0.18)	0.00 (0.13)	99.65%	-0.01 (0.11)
	K_2	-0.03 (0.20)	0 (0)	0.02 (0.19)	0.01 (0.13)	99.58%	-0.01 (0.12)
1000	K_1	-0.01 (0.14)	0 (0)	0.01 (0.13)	0.00 (0.09)	99.79%	-0.01 (0.08)
	K_2	0.01 (0.13)	0 (0)	0.00 (0.13)	0.00 (0.10)	99.79%	-0.01 (0.08)

satisfactory performance.

Example A4 (addition results fro real-data example). Table A.5 shows the smooth and nonsmooth estimators for the real example in Section 2.5, with a 5-fold cross-validation. Specifically, we randomly divide the data into five folds and use four folds to estimate $\boldsymbol{\beta}_0$ and $V(\beta_0)$ and evaluate the matching ratio on the remaining fold (i.e., validation data). Each of the five folds is used as validation data in turn (refereed to as iterative 1, . . . , 5 in te table), the final results are summarized in Table A.5. We observe that the smooth estimators always lie in the element-wise confidence intervals we calculated in Section 2.5. However, the nonsmooth estimators are rather nonstable and can even change signs across iteratives.

Table A.4: Monte Carlo estimates of the bias and standard deviation of the estimate for the parameters indexing the optimal treatment regime, the match ratio (percentage of times the estimated optimal treatment regime matches the theoretically optimal treatment regime), and the bias and standard deviation of the estimated optimal value in an observational study.

n	β_0^{opt}	β_1^{opt}	β_2^{opt}	β_3^{opt}	Match Ratio	$V_n(\widehat{\boldsymbol{\beta}}_n)$
			Setting 1			
300	-0.05 (0.28)	0 (0)	0.03 (0.32)	0.04 (0.33)	99.40%	-0.01 (0.15)
500	-0.03 (0.23)	0 (0)	0.03 (0.29)	0.03 (0.27)	99.63%	-0.01 (0.12)
1000	-0.03 (0.15)	0 (0)	0.03 (0.18)	0.03 (0.18)	99.59%	-0.01 (0.08)
			Setting 2			
300	-0.06 (0.27)	0 (0)	0.03 (0.29)	0.01 (0.19)	99.01%	-0.01 (0.15)
500	-0.02 (0.20)	0 (0)	0.02 (0.23)	0.01 (0.15)	99.66%	-0.01 (0.12)
1000	-0.02 (0.14)	0 (0)	0.02 (0.15)	0.01 (0.10)	99.69%	-0.01 (0.09)

Table A.5: Real data example: comparison of smooth and nonsmooth estimators based on 5-fold cross-validation

Iterative	Method	β_0^{opt}	β_1^{opt}	β_2^{opt}	Match Ratio
1	Smooth	0.65	1	0.46	87.93%
	Nonsmooth	-0.40	1	0.36	
2	Smooth	0.45	1	0.35	86.21%
	Nonsmooth	-0.39	1	-0.35	
3	Smooth	0.64	1	0.42	89.47%
	Nonsmooth	-0.42	1	-0.16	
4	Smooth	0.46	1	0.27	85.96%
	Nonsmooth	0.14	1	-0.73	
5	Smooth	0.64	1	0.46	87.72%
	Nonsmooth	-0.40	1	13.38	

Appendix B

Supporting Information for Chapter 3

B.1 Chapter Outline

The chapter is constructed as follows. Appendix B.2 states some regularity conditions and useful lemmas for theorems in Chapter 3. Appendix B.3 and Appendix B.4 of the supplementary material provide the proofs of the theoretical results in Section 3.3.1 and Section 3.3.2 of Chapter 3, respectively. Appendix B.5 presents the proofs of the technical lemmas in Appendix B.2. Appendix B.6 provides additional technical details.

B.2 Regularity Conditions and Some Technical Lemmas

We first state some regularity conditions. (B1) includes the assumptions about data and model (3.1); (B2) are assumptions about $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})$; (B3) states assumptions about the kernel function $K(\cdot)$; (B4) states assumptions about the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$, and (B5) involves assumptions about function $G(t|\boldsymbol{\beta})$.

(B1) $\boldsymbol{x} \in \mathbb{R}^P$ and ϵ are both sub-Gaussian with variance proxy σ_x^2 and σ_ϵ^2, respectively. $||\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)||_\infty \leqslant M_1$ for constant $M_1 > 0$, and the smallest eigenvalue of $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)$, denoted as ξ_1, is strictly positive and $\xi_1^{-1} \leqslant M_2$ for constant $M_2 > 0$. $P(|g(\boldsymbol{x})| \leqslant M_g) = 1$. $f_0(z)$ is twice differentiable and increasing for $z \in \mathbb{R}$, with $f_0(0) = 0$; and

$0 < a \leqslant f_0'(z) \leqslant b$ for any $z \in \mathbb{R}$; $f_0''(z)$ exists and is bounded for $z \in \mathbb{R}$.

(B2) The smallest eigenvalue of $\mathrm{E}\big[\,\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]$, denoted as ξ_0, is strictly positive. The largest eigenvalue of $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{x}^T\boldsymbol{\beta}_0)$ is bounded almost surely. For any $\boldsymbol{\beta} \in \mathbb{B}$, $||\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)||_\infty \leqslant L|\boldsymbol{x}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|$, and $||\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{x}^T\boldsymbol{\beta})-\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{x}^T\boldsymbol{\beta}_0)||_\infty \leqslant L|\boldsymbol{x}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)|$ almost surely with constant $L > 0$.

(B3) $K(\cdot)$ is twice differentiable and bounded on the real line. $K(\cdot)$, $K'(\cdot)$ and $K''(\cdot)$ are all Lipschitz on the real line. $K(\cdot)$ is nonnegative and symmetric about 0. Furthermore, $\lim_{|\nu|\to\infty} K(\nu) = 0$, $\int_{-\infty}^{\infty} K(\nu)d\nu = 1$, $\int_{-\infty}^{\infty} \nu K'(\nu)d\nu = -1$, and $\int_{-\infty}^{\infty} \nu^2 K''(\nu)d\nu = 2$. For any integer $0 \leqslant i \leqslant 4$ $\int |\nu^i K(\nu)|d\nu < \infty$; for integer $0 \leqslant i \leqslant 3$, $\int |\nu^i K'(\nu)|d\nu < \infty$; for integer $0 \leqslant i \leqslant 2$, $\int |\nu^i K''(\nu)|d\nu < \infty$.

(B4) Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. $f_{\boldsymbol{\beta}}(\cdot)$ is twice differentiable. $f_{\boldsymbol{\beta}}(\cdot)$, $f'_{\boldsymbol{\beta}}(\cdot)$ and $f''_{\boldsymbol{\beta}}(\cdot)$ are all bounded on the real line; and $f_{\boldsymbol{\beta}}^{-1}(t) \leqslant M_3$ for positive constant M_3, $t \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{B}$.

(B5) (a) For any $\boldsymbol{\beta} \in \mathbb{B}$ and $t \in \mathbb{R}$, $G(t|\boldsymbol{\beta})$ is twice differentiable with respect to t; $G^{(2)}(t|\boldsymbol{\beta})$ is bounded.

(b) $\inf_{\boldsymbol{\beta}\in\mathbb{B}} \inf_{|t|\leqslant \log p} G^{(1)}(t|\boldsymbol{\beta}) \geqslant a$, and $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \sup_{|t|\leqslant \log p} G^{(1)}(t|\boldsymbol{\beta}) \leqslant b$, for positive constants $a \leqslant b$.

(c) $G(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ and $G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ both satisfy Lipschitz condition as follows:

$$\begin{aligned}\big|G(\boldsymbol{x}^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1) - G(\boldsymbol{x}^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\big| &\leqslant L|\boldsymbol{x}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)|;\\ \big|G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_1|\boldsymbol{\beta}_1) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_2|\boldsymbol{\beta}_2)\big| &\leqslant L|\boldsymbol{x}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)|,\end{aligned}$$

with positive constant L, for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{B}$ and $\boldsymbol{x} \in \mathbb{R}^p$.

Next, we state several useful lemmas in the proofs of theorems and corollaries. Their proofs can be found in Appendix B.5 of the online supplementary material.

Lemma B.1

Under assumptions of Theorem 3.1, there exist universal positive constants d_0 and d_1 such that $\left|\left|\boldsymbol{S}_n(\boldsymbol{\beta}_0)\right|\right|_\infty \leqslant d_0\sqrt{\frac{\log p}{n}}$ with probability at least $1-\exp(-d_1\log p)$. □

Lemma B.2

If $\boldsymbol{x} \in \mathbb{R}^p$ is sub-Gaussian with variance proxy σ^2, then $\mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)$ and $\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)$ are both sub-Gaussian, with variance proxy σ^2 and $2\sigma^2$, respectively. Furthermore, under assumption (B1), $\widetilde{\epsilon} = 2(2A-1)\big(\epsilon + g(\boldsymbol{x})\big)$ is also sub-Gaussian with variance proxy $16(\sigma_\epsilon^2 + M_g^2)$. □

Lemma B.3

Under assumption (B1) and (B5), let us define the following three events:

$$\begin{aligned}
\mathcal{G}_n &= \Big\{ \max_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} \big|G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j\big| \leqslant \sigma_x\sqrt{\log(p\vee n)}\Big\},\\
\mathcal{H}_n &= \Big\{ \max_{1\leqslant j\leqslant p, 1\leqslant i\leqslant n} \big|\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j\big| \leqslant \sigma_x(\sigma_\epsilon + M_g)\log(p\vee n)\Big\},\\
\mathcal{K}_n &= \{\max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty \leqslant \sigma_x\sqrt{\log(p\vee n)}\}.
\end{aligned}$$

There is a universal positive constant c such that all the events hold with probability at least $1-\exp[-c\log(p\vee n)]$, for sufficiently large n. □

Lemma B.4

Under assumptions of Theorem 3.1, denote $\mathbb{T} = \{t\in\mathbb{R} : |t| \leqslant \sqrt{\log(p\vee n)}\}$, there exist universal positive constants c_0 and c_1 such that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} \big|\widehat{G}(t|\boldsymbol{\beta}) - G(t|\boldsymbol{\beta})\big| \geqslant c_0h^2\Big) \leqslant \exp(-c_1nh^5).$$ □

Lemma B.5

Under assumptions of Theorem 3.1, there exist universal positive constants c_0 and c_1 such that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}\big|\widehat{G}^{(1)}(t|\boldsymbol{\beta})-G^{(1)}(t|\boldsymbol{\beta})\big|\geqslant c_0h\Big)\leqslant\exp(-c_1nh^5). \qquad \square$$

Lemma B.6

Let $\mathbb{B}_1=\big\{\boldsymbol{\beta}\in\mathbb{B}:||\boldsymbol{\beta}-\boldsymbol{\beta}_0||_1\leqslant c_0s\max\big\{h^2,\sqrt{\frac{\log p}{n}}\big\},||\boldsymbol{\beta}||_0\leqslant ks\big\}$ for positive constants c_0 and k, with $ks=O(n)$. Under assumption (B1) and $s\max\big\{h^2,\sqrt{\frac{\log p}{n}}\big\}\leqslant d_0$ for positive constant d_0 and all sufficiently large n, there exist universal positive constants c_0 and c_1 such that

$$P\big(\max_{1\leqslant i\leqslant n}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}|\boldsymbol{x}_i^T\boldsymbol{\beta}|>c_0\sqrt{\log(p\vee n)}\big)\leqslant\exp[-c_1\log(p\vee n)].$$

Lemma B.7

Under assumptions of Theorem 3.2, let $\gamma_1=\sqrt{h\log(p\vee n)}+h[\log(p\vee n)]^{3/2}$, for constant $c>0$,

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\big|n^{-1/2}\sum_{i=1}^n\boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z_i,\boldsymbol{\beta},m)\big|>\gamma\Big)\leqslant\exp-[c\log(p\vee n)], \qquad \text{(B.1)}$$

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\big|\big|n^{-1/2}\sum_{i=1}^n\boldsymbol{\gamma}(Z_i,\boldsymbol{\beta},m)\big|\big|_\infty>\gamma\Big)\leqslant\exp-[c\log(p\vee n)], \qquad \text{(B.2)}$$

$\square$

where $Z_i=(\boldsymbol{x}_i,\epsilon_i,A_i)$, $\boldsymbol{\gamma}(Z_i,\boldsymbol{\beta},m)=\big[m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\widetilde{\epsilon}_i\boldsymbol{x}_i$, $||m||_\infty=\sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||m(\cdot|\boldsymbol{\beta})||_\infty$, $m(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and

$$\mathbb{M}=\Big\{m(\cdot|\boldsymbol{\beta}):\boldsymbol{\beta}\in\mathbb{B}_1,\ m(\cdot|\boldsymbol{\beta})\text{ is Lipschitz }\forall\boldsymbol{\beta},\ \sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||m(\cdot|\boldsymbol{\beta})-G^{(1)}(\cdot|\boldsymbol{\beta}_0)||_\infty\leqslant c_1h\Big\}.$$

Lemma B.8

Under assumptions of Theorem 3.2, for positive constant c,

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_1(Z_i,\boldsymbol{\beta},u)\big|>h^{3/4}\Big)\leqslant\exp-[c\log(p\vee n)],\tag{B.3}$$

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,u\in\mathbb{U}}\big|\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\nu}_1(Z_i,\boldsymbol{\beta},u)\big|\big|_\infty>h^{3/4}\Big)\leqslant\exp-[c\log(p\vee n)],\tag{B.4}$$

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)\big|>h\Big)\leqslant\exp-[c\log(p\vee n)],\tag{B.5}$$

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{\boldsymbol{\beta}\in\mathbb{B}_1,m\in\mathbb{M}}\big|\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)\big|\big|_\infty>h\Big)\leqslant\exp-[c\log(p\vee n)],\tag{B.6}$$

□

where $\boldsymbol{\nu}_1(Z_i,\boldsymbol{\beta},u)=\big[u(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-u(\boldsymbol{x}_i^T\boldsymbol{\beta})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i$, $\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)=m(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i$, $u(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{\beta})$ depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and

$$\mathbb{U}=\Big\{u(\cdot|\boldsymbol{\beta}):\boldsymbol{\beta}\in\mathbb{B}_1,\ u(\cdot|\boldsymbol{\beta})\text{ is Lipschitz }\forall\boldsymbol{\beta},\ \sup_{\boldsymbol{\beta}\in\mathbb{B}_1}||u(\cdot|\boldsymbol{\beta})-G(\cdot|\boldsymbol{\beta}_0)||_\infty\leqslant c_1h^2\Big\}.$$

Lemma B.9

Under assumptions of Theorem 3.1, for positive constant c,

$$P\Big(\max_{1\leqslant j\leqslant p}\sup_{u\in\mathbb{U}}\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|>c_0h^{3/4}\Big)\leqslant\exp[-c_1\log(p\vee n)],\tag{B.7}$$

$$P\Big(\sup_{u\in\mathbb{U}}\big|\big|n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|\big|_\infty>c_0h^{3/4}\Big)\leqslant\exp[-c_1\log(p\vee n)],\tag{B.8}$$

where $\boldsymbol{\psi}(\boldsymbol{x}_i,u)=\big[u(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i$, $u(\cdot|\boldsymbol{\beta})$ and $\mathbb{U}$ follow the definitions in Lemma B.8. □

Lemma B.10

Under assumptions of Lemma 3.2, there are universal positive constants c_0 and c_1, and

$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^p : ||\boldsymbol{x}||_\infty \leqslant \sqrt{\log(p \vee n)}\}$, such that

$$P\Big(\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1} \big|\widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big| \geqslant c_0 s h^2 \sqrt{\log(p \vee n)} \Big) \leqslant \exp(-c_1 n h^5), \quad \text{(B.9)}$$

$$P\Big(\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1} \big|\widehat{G}^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big| \geqslant c_0 h \Big) \leqslant \exp(-c_1 n h^5). \quad \text{(B.10)}$$

□

Lemma B.11

Under assumptions (B1) and (B5), we have

(1) $\inf_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v} \geqslant a^2\xi_1$; (2) $\tau_{0j}^{-2} \leqslant ||\boldsymbol{\theta}_j||_2 \leqslant M_2/a^2$ and $\tau_{0j}^2 \leqslant b^2 M_1$ uniformly in j. □

Lemma B.12

Under assumptions of Theorem 3.2, $||\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}||_\infty = o_p(1)$. □

B.3 Proofs of results in Section 3.3.1

Proof of Lemma 3.1 Observe

$$\begin{aligned}
\boldsymbol{S}_n(\boldsymbol{\beta}) &= -n^{-1}\sum_{i=1}^n \big\{\widetilde{\epsilon}_i + G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta})\big\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i, \\
\boldsymbol{S}_n(\boldsymbol{\beta}_0) &= -n^{-1}\sum_{i=1}^n \big\{\widetilde{\epsilon}_i + G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i.
\end{aligned}$$

Denote $\boldsymbol{\eta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$. We have

$$
\begin{aligned}
& \big\langle \boldsymbol{S}_n(\boldsymbol{\beta}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0), \boldsymbol{\eta} \big\rangle \\
= \; & -n^{-1} \sum_{i=1}^{n} \widetilde{\epsilon}_i \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} + n^{-1} \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& + n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}) \big] \widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x}_i^T \boldsymbol{\eta} - n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \boldsymbol{x}_i^T \boldsymbol{\eta} \\
= \; & n^{-1} \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x}_i^T \boldsymbol{\eta} - n^{-1} \sum_{i=1}^{n} \widetilde{\epsilon}_i \big[G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& + n^{-1} \sum_{i=1}^{n} \big[G(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& - n^{-1} \sum_{i=1}^{n} \widetilde{\epsilon}_i \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - \big(\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& + n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}) \big] G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& + n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}) \big] \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& - n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \boldsymbol{x}_i^T \boldsymbol{\eta} \\
& - n^{-1} \sum_{i=1}^{n} \big[\widehat{G}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) \big] \boldsymbol{x}_i^T \boldsymbol{\eta} \\
= \; & \sum_{l=1}^{8} J_{nl},
\end{aligned}
$$

where J_{nl} ($l = 1, \cdots, 8$) are defined clearly from the context.

By Taylor expansion,

$$
\begin{aligned}
& G(\boldsymbol{x}_i^T\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) = \mathrm{E}\big[f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)|\boldsymbol{x}_i^T\boldsymbol{\beta}\big] - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) \\
= \;& \mathrm{E}\big[f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}) - f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i^T\boldsymbol{\eta} + \frac{1}{2}f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big] - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) \\
= \;& f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}) - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T\boldsymbol{\eta} + \frac{1}{2}\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big] \\
= \;& f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i^T\boldsymbol{\eta} + \frac{1}{2}f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2 - f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})^T\boldsymbol{\eta} + \frac{1}{2}\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big] \\
= \;& f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]^T\boldsymbol{\eta} + \frac{1}{2}f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2 + \frac{1}{2}\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big],
\end{aligned}
$$

for some $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. We therefore write $J_{n1} = n^{-1}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i^T\boldsymbol{\eta} = \sum_{k=1}^5 J_{n1k}$, where

$$
\begin{aligned}
J_{n11} &= n^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\eta}^T\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T\boldsymbol{\eta}, \\
J_{n12} &= n^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\eta}^T\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)^T\boldsymbol{\eta}, \\
J_{n13} &= n^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})f_0'(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{\eta}^T\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\boldsymbol{x}_i^T\boldsymbol{\eta}, \\
J_{n14} &= (2n)^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_2)(\boldsymbol{x}_i^T\boldsymbol{\eta})^3, \\
J_{n15} &= (2n)^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta})\mathrm{E}\big[f_0''(\boldsymbol{x}_i^T\boldsymbol{\beta}_1)(\boldsymbol{x}_i^T\boldsymbol{\eta})^2|\boldsymbol{x}_i^T\boldsymbol{\beta}\big]\boldsymbol{x}_i^T\boldsymbol{\eta}.
\end{aligned}
$$

Assumption (B2) implies that

$$
J_{n11} \geqslant a^2\xi_0||\boldsymbol{\eta}||_2^2 - b^2\Big|n^{-1}\sum_{i=1}^n \boldsymbol{\eta}^T\big\{\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T - \mathrm{E}\big[\,\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\big\}\boldsymbol{\eta}\Big|.
$$

Note that $\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ is sub-Gaussian with covariance matrix $\mathrm{E}\big[\,\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]$. Denote $\boldsymbol{\Psi}_n = n^{-1}\sum_{i=1}^n \big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\big[\boldsymbol{x}_i - \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]^T - \mathrm{E}\big[\,\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]$.

Lemma B.14 implies that

$$P\Big(\sup_{\boldsymbol{v}\in\mathbb{K}(ks)}\Big|\boldsymbol{v}^T\boldsymbol{\Psi}_n\boldsymbol{v}\Big|\geqslant\sigma_x^2\sqrt{\frac{s\log p}{n}}\Big)\leqslant\exp(-c_1 s\log p),$$

for a universal positive constant c_1. Hence with probability at least $1-\exp(-c_1 s\log p)$, we have $J_{n11}\geqslant||\boldsymbol{\eta}||_2^2\big(a^2\xi_0-\sqrt{\frac{s\log p}{n}}\big)$ for any $\boldsymbol{\eta}=\boldsymbol{\beta}-\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}\in\mathbb{B}$.

Similarly for J_{n12}, $\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ and $\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ are both sub-Gaussian, with variance proxy $2\sigma_x^2$ and σ_x^2, respectively. Note that $\mathrm{Cov}\left\{\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0),\ \mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right\}=\mathbf{0}_{p\times p}$. Hence, Lemma B.13 implies

$$P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\big[\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\Big|\Big|_\infty\geqslant c_0\sigma_x^2\sqrt{\frac{\log p}{n}}\Big)\leqslant\exp(-c_1\log p).$$

We thus have

$$|J_{n12}|\leqslant\frac{b^2}{n}\Big|\sum_{i=1}^n\boldsymbol{\eta}^T\big[\boldsymbol{x}_i-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)^T\boldsymbol{\eta}\Big|\leqslant c_0||\boldsymbol{\eta}||_1^2\sqrt{\frac{\log p}{n}},$$

with probability at least $1-\exp(-c_1\log p)$ for universal positive constants c_0 and c_1. To evaluate J_{n13}, we observe that

$$\begin{aligned}|J_{n13}| &\leqslant \frac{b^2}{n}\sum_{i=1}^n\Big|\boldsymbol{\eta}^T\big[\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big]\boldsymbol{x}_i^T\boldsymbol{\eta}\Big|\\ &\leqslant \frac{b^2}{n}\sum_{i=1}^n||\boldsymbol{\eta}||_1\big|\big|\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\mathrm{E}(\boldsymbol{x}_i|\boldsymbol{x}_i^T\boldsymbol{\beta})\big|\big|_\infty|\boldsymbol{x}_i^T\boldsymbol{\eta}|\leqslant\frac{b^2Lr}{n}\sum_{i=1}^n(\boldsymbol{x}_i^T\boldsymbol{\eta})^2.\end{aligned}$$

Since $\boldsymbol{x}_i$ is mean-zero sub-Gaussian with variance proxy σ_x^2, similarly as previous steps, we have that

$$\frac{1}{n}\sum_{i=1}^n(\boldsymbol{x}_i^T\boldsymbol{\eta})^2\leqslant\Big(\xi_p+c\sqrt{\frac{s\log p}{n}}\Big)||\boldsymbol{\eta}||_2^2,$$

with probability at least $1-\exp(-c_1 s\log p)$, for a universal constant $c>0$ and any any

$\boldsymbol{\eta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and $\boldsymbol{\beta} \in \mathbb{B}$. It follows that $|J_{n13}| \leqslant c_0 r||\boldsymbol{\eta}||_2^2$. Lemma B.15 ensures that $|J_{n14}| \leqslant c_0\sigma_x^3||\boldsymbol{\eta}||_2^3$ with probability at least $1 - \exp\big[-c_1 \min\{s\log p, (ns\log p)^{1/3}\}\big]$.

To bound $|J_{n15}|$, assumption (B2) implies that

$$\begin{aligned} |J_{n15}| &\leqslant \frac{c}{n}\sum_{i=1}^n \Big|\boldsymbol{\eta}^T \mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{\eta}\boldsymbol{x}_i^T\boldsymbol{\eta}\Big| + \frac{c}{n}\sum_{i=1}^n \Big|\boldsymbol{\eta}^T\big[\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta}) - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T|\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\boldsymbol{\eta}\boldsymbol{x}_i^T\boldsymbol{\eta}\Big| \\ &\leqslant \frac{c||\boldsymbol{\eta}||_2^2}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\eta}| + \frac{c||\boldsymbol{\eta}||_1^2}{n}\sum_{i=1}^n (\boldsymbol{x}_i^T\boldsymbol{\eta})^2. \end{aligned}$$

Note that $\frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\eta}| \leqslant \sqrt{\frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\eta}|^2} \leqslant c_0||\boldsymbol{\eta}||_2$ for any $\boldsymbol{\eta}$, with probability at least $1 - \exp(-c_1 s\log p)$. Since $s_0 \geqslant 1$, $|J_{n15}| \leqslant c_0 r||\boldsymbol{\eta}||_2^2$ with the same probability bound.

Combining all the preceding results, we conclude that $J_{n1} \geqslant c_0\big(||\boldsymbol{\eta}||_2^2 - ||\boldsymbol{\eta}||_1^2\sqrt{\frac{\log p}{n}}\big)$ with probability at least $1 - \exp\big\{-c_1[(ns\log p)^{1/3} \wedge \log p]\big\}$, with $0 < r < 1$ sufficiently small, and universal positive constants c_0 and c_1.

Note that $\sup_{\boldsymbol{\beta}\in\mathbb{B}} \max_{1\leqslant i\leqslant n} |\boldsymbol{x}_i^T\boldsymbol{\beta}| \leqslant \sup_{\boldsymbol{\beta}\in\mathbb{B}} ||\boldsymbol{\beta}||_1 * \max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty$. By the sub-Gaussian property for $\boldsymbol{x}_i$, and consider $||\boldsymbol{\beta}_0||_1$ as a constant and r sufficiently small, we have

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}}\max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T\boldsymbol{\beta}| \geqslant \sqrt{\log(p\vee n)}\Big) \leqslant P\Big(\max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty \geqslant \frac{\sqrt{\log(p\vee n)}}{||\boldsymbol{\beta}_0||_1 + r}\Big) \leqslant \exp[-c_1\log(p\vee n)],$$

for universal constant $c_1 > 0$. Conditional on the event $\{\sup_{\boldsymbol{\beta}\in\mathbb{B}} \max_{1\leqslant i\leqslant n} |\boldsymbol{x}_i^T\boldsymbol{\beta}| \leqslant \sqrt{\log(p\vee n)}\}$, Lemma B.4 implies that with probability at least $1 - \exp[-c_2\log(p\vee n)]$,

$$|J_{n5}| \leqslant b \sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} \big|\widehat{G}(t|\boldsymbol{\beta}) - G(t|\boldsymbol{\beta})\big| * n^{-1}\sum_{i=1}^n |\boldsymbol{x}_i^T\boldsymbol{\eta}| \leqslant d_1 h^2||\boldsymbol{\eta}||_2,$$

holds for some universal constant $d_1 > 0$. Similarly, we can also show that for universal

constant $d_1 > 0$,

$$|J_{n2}| \leqslant d_1||\boldsymbol{\eta}||_2\sqrt{\frac{\log p}{n}}, \qquad |J_{n3}| \leqslant d_1 h||\boldsymbol{\eta}||_2^2, \qquad |J_{n4}| \leqslant d_1 h||\boldsymbol{\eta}||_2\sqrt{\frac{\log p}{n}},$$
$$|J_{n6}| \leqslant d_1 h^3||\boldsymbol{\eta}||_2, \qquad |J_{n7}| \leqslant d_1 h^2||\boldsymbol{\eta}||_2, \qquad |J_{n8}| \leqslant d_1 h^3||\boldsymbol{\eta}||_2,$$

hold with probability at least $1 - \exp(-c_1 \log p)$. Since $n^{-1}\log p = O(h^5)$, there exist universal positive constants c_0, c_1, c_2 and r such that

$$\big\langle \boldsymbol{S}_n(\boldsymbol{\beta}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0), \boldsymbol{\eta}\big\rangle \geqslant c_0||\boldsymbol{\eta}||_2^2 - c_1 h^2||\boldsymbol{\eta}||_1,$$

holds for any $\boldsymbol{\beta} \in \mathbb{B}$ with probability at least $1 - \exp\big\{-c_1[(ns\log p)^{1/3} \wedge \log p]\big\}$. □

Proof of Theorem 3.1 By the definition of $\widehat{\boldsymbol{\beta}}$, we have

$$\big\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\big\rangle = 0, \tag{B.11}$$

for all feasible $\boldsymbol{\beta}$, where $\widehat{\boldsymbol{\kappa}} \in \partial||\widehat{\boldsymbol{\beta}}||_1$. In particular, $\big\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\big\rangle = 0$. By the property of convex function, we know that $||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1 \geqslant \big\langle \widehat{\boldsymbol{\kappa}}, \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\big\rangle$, for any $\widehat{\boldsymbol{\kappa}} \in \partial||\widehat{\boldsymbol{\beta}}||_1$. Putting these two results together, we have

$$\big\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\eta}}\big\rangle = \lambda\big\langle \widehat{\boldsymbol{\kappa}}, -\widehat{\boldsymbol{\eta}}\big\rangle \leqslant \lambda(||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1).$$

Let $\widehat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and apply the local restricted strong convexity condition established in Lemma 3.1 to $\big\langle \boldsymbol{S}_n(\boldsymbol{\beta}) - \boldsymbol{S}_n(\boldsymbol{\beta}_0), \boldsymbol{\eta}\big\rangle$. We obtain

$$\begin{aligned} c_0||\widehat{\boldsymbol{\eta}}||_2^2 - c_1 h^2||\widehat{\boldsymbol{\eta}}||_1 &\leqslant \big\langle \boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\eta}}\big\rangle - \big\langle \boldsymbol{S}_n(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\eta}}\big\rangle \\ &\leqslant \lambda(||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1) - \big\langle \boldsymbol{S}_n(\boldsymbol{\beta}_0), \widehat{\boldsymbol{\eta}}\big\rangle \\ &\leqslant \lambda(||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1) + ||\boldsymbol{S}_n(\boldsymbol{\beta}_0)||_\infty||\widehat{\boldsymbol{\eta}}||_1, \end{aligned} \tag{B.12}$$

with probability at least $1-\exp(-c_1 \log p)$. It implies that $c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant (c_1 h^2 + ||\boldsymbol{S}_n(\boldsymbol{\beta}_0)||_\infty)||\widehat{\boldsymbol{\eta}}||_1 + \lambda(||\boldsymbol{\beta}_0||_1 - ||\widehat{\boldsymbol{\beta}}||_1)$. Lemma B.1 implies that we can take λ such that $\lambda/4 \geqslant c_1 h^2$, and $\lambda/4 \geqslant ||\boldsymbol{S}_n(\boldsymbol{\beta}_0)||_\infty$ with probability at least $1 - \exp(-c_1 \log p)$. Then we have

$$c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant \frac{\lambda}{2}(||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 + ||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1) + \lambda((||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 - ||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1) \leqslant \frac{3\lambda}{2}||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 - \frac{\lambda}{2}||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1,$$

which implies that $||\widehat{\boldsymbol{\eta}}_{\mathcal{S}^C}||_1 \leqslant 3||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1$. Then (B.12) implies that

$$c_0||\widehat{\boldsymbol{\eta}}||_2^2 \leqslant (c_1 h^2 + ||\nabla L_n(\boldsymbol{\beta}_0)||_\infty + \lambda)||\widehat{\boldsymbol{\eta}}||_1 \leqslant \frac{3\lambda}{2}||\widehat{\boldsymbol{\eta}}||_1 \leqslant 6\lambda||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 \leqslant 6\lambda\sqrt{s}||\widehat{\boldsymbol{\eta}}||_2.$$

Hence $||\widehat{\boldsymbol{\eta}}||_2 \leqslant \frac{6}{c_0}\lambda\sqrt{s}$. Since $||\widehat{\boldsymbol{\eta}}||_1 \leqslant 4||\widehat{\boldsymbol{\eta}}_{\mathcal{S}}||_1 \leqslant 4\sqrt{s}||\widehat{\boldsymbol{\eta}}||_2$, the bound on $||\widehat{\boldsymbol{\eta}}||_1$ follows immediately. □

B.4 Proofs of results in Section 3.3.2

Proof of Lemma 3.2 (1) To derive the uniform error bound for $\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}}, \eta)$, we first prove the following two results:

(a) $\boldsymbol{d}_{0j}$ is feasible for the constraint of Dantzig Selector problem with high probability, uniformly in $j = 1, \cdots, p$;

(b) $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 \boldsymbol{x}_i \boldsymbol{x}_i^T$ satisfies the restricted eigenvalue condition on the support of $\boldsymbol{d}_{0j}$, denoted by $\mathcal{S}_{d_j}$, with high probability, uniformly in $j = 1, \cdots, p$.

To prove (a), let $\boldsymbol{\phi}_{0j} = \tau_{0j}^2 \boldsymbol{\theta}_j$, then $\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T \boldsymbol{d}_{0j} = \boldsymbol{x}_i^T \boldsymbol{\phi}_{0j}$. Note that $\sigma_x^2\sqrt{\frac{\log p}{n}} = o(h^{5/2}) = o(\eta)$. Lemma B.19 implies that

$$P\Big(\max_{1\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^n [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2 \boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant \eta/2\Big) \leqslant \exp(-c\log p).$$

Lemma B.20 implies $P\Big(\max_{1\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^n \{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant$

$c_0 h\Big) \leqslant \exp(-c_1 \log p)$. Hence there exists a universal constant $c_1 > 0$, such that

$$\begin{aligned}
& P\Big(\max_{1\leqslant j\leqslant p} \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2(\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T\boldsymbol{d}_{0j})\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant \eta\Big) \\
\leqslant\ & P\Big(\max_{1\leqslant j\leqslant p} \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2(\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T\boldsymbol{d}_{0j})\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant \eta/2\Big) \\
& + P\Big(\max_{1\leqslant j\leqslant p} \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}(\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T\boldsymbol{d}_{0j})\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant \eta/2\Big) \\
\leqslant\ & \exp(-c_1 \log p).
\end{aligned}$$

This ensures that $\boldsymbol{d}_{0j}$ is feasible for the Dantzig Selector problem with high probability uniformly in j. It follows that $P\Big[\max_{1\leqslant j\leqslant p}\big(||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)||_1 - ||\boldsymbol{d}_{0j}||_1\big) > 0\Big] \leqslant \exp(-c_1\log p)$.

To prove (b), consider the set $\mathcal{C}(\mathcal{I}) = \{\boldsymbol{v}\in\mathbb{R}^p : ||\boldsymbol{v}_{\mathcal{I}^C}||_1 \leqslant ||\boldsymbol{v}_{\mathcal{I}}||_1, ||\boldsymbol{v}||_2 = 1\}$ for any support $\mathcal{I}$. Lemma B.11 indicates that

$$\inf_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \boldsymbol{v}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{v} \geqslant \inf_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v} - \sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\}\boldsymbol{v} \geqslant a^2\xi_1 - \sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\}\boldsymbol{v}.$$

Note that

$$\begin{aligned}
\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \big|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\}\boldsymbol{v}\big| \ \leqslant\ & \sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \Big|\frac{1}{n}\sum_{i=1}^{n}\Big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\Big\}(\boldsymbol{x}_i^T\boldsymbol{v})^2\Big| \\
& + \sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})} \Big|\boldsymbol{v}^T\Big(\frac{1}{n}\sum_{i=1}^{n}\Big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\Big\}\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{\Omega}\Big)\boldsymbol{v}\Big|.
\end{aligned}$$

Lemma B.5 and the proof of Lemma 3.1 imply that with probability at least $1-\exp(-c_1\log p)$, $\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\Big\}(\boldsymbol{x}_i^T\boldsymbol{v})^2\Big| \leqslant c_0 h$, for universal positive constants c_0 and c_1. Theorem 2.1 in Wainwright (2015) implies that $(2A_i - 1)G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i$ is

sub-Gaussian with variance proxy no larger than $b^2\sigma_x^2$. Lemma B.14 indicates

$$P\Big(\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})}\Big|\boldsymbol{v}^T\Big(\frac{1}{n}\sum_{i=1}^n\Big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\Big\}\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{\Omega}\Big)\boldsymbol{v}\Big|\geqslant c_0\sigma_x^2\sqrt{\frac{|\mathcal{I}|\log p}{n}}\Big)\leqslant\exp(-c_2|\mathcal{I}|\log p).$$

Hence we have

$$P\Big(\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})}\big|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\big|\geqslant c_0\big(h+\sqrt{n^{-1}|\mathcal{I}|\log p}\big)\Big)\leqslant\exp(-c_1\log p).$$

Note that $\sqrt{\frac{\widetilde{s}\log p}{n}}=o(\sqrt{\widetilde{s}h^5})=o(h)$. Hence we conclude

$$\begin{aligned}P\Big(\min_{1\leqslant j\leqslant p}\inf_{\boldsymbol{v}\in\mathcal{C}(\mathcal{S}_{d_j})}\boldsymbol{v}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{v}\leqslant\frac{a^2\xi_1}{2}\Big)&\leqslant P\Big(a^2\xi_1-\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})}\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\leqslant\frac{a^2\xi_1}{2}\Big)\\&\leqslant P\Big(\sup_{\boldsymbol{v}\in\mathcal{C}(\mathcal{I})}\big|\boldsymbol{v}^T\{\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega}\}\boldsymbol{v}\big|\geqslant\frac{a^2\xi_1}{2}\Big)\leqslant\exp(-c_1\log p).\end{aligned}$$

It proves (b).

Given (a) and (b), the event

$$\mathcal{E}=\Big\{\min_{1\leqslant j\leqslant p}\inf_{\boldsymbol{v}\in\mathcal{C}(\mathcal{S}_{d_j})}\boldsymbol{v}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{v}\geqslant\frac{a^2\xi_1}{2},\text{ and }\boldsymbol{d}_{0j}\text{ is feasible uniformly in }j\Big\},$$

holds with probability at least $1-\exp(-c_1\log p)$. Define $\boldsymbol{w}_j=\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}},\eta)-\boldsymbol{\phi}_{0j}$, then for any j, conditional on $\mathcal{E}$,

$$||\boldsymbol{w}_{j,\mathcal{S}_{d_j}^C}||_1=\big|\big|\big[\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)\big]_{\mathcal{S}_{d_j}^C}\big|\big|_1\leqslant||\boldsymbol{d}_{0j,\mathcal{S}_{d_j}}||_1-\big|\big|\big[\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)\big]_{\mathcal{S}_{d_j}}\big|\big|_1\leqslant||\boldsymbol{w}_{j,\mathcal{S}_{d_j}}||_1,$$

which means that $\boldsymbol{w}_j \in \mathcal{C}(\mathcal{S}_{d_j})$. In addition,

$$\begin{aligned} ||\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j||_\infty &\leqslant \Big|\Big|\frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\big(\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta)\big)\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \\ &\quad + \Big|\Big|\frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\big(\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,-j}^T\boldsymbol{d}_{0j}\big)\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \leqslant 2\eta. \end{aligned}$$

Hence we conclude that

$$s_j^{-2}||\boldsymbol{w}_j||_1^2 \leqslant 4s_j^{-1}||\boldsymbol{w}_j||_2^2 \leqslant \frac{8\boldsymbol{w}_j^T\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j}{s_j a^2\xi_1} \leqslant \frac{8||\boldsymbol{w}_j||_1||\widehat{\boldsymbol{\Omega}}\boldsymbol{w}_j||_\infty}{s_j a^2\xi_1} \leqslant \frac{16\eta||\boldsymbol{w}_j||_1}{s_j a^2\xi_1} \leqslant \frac{32\eta||\boldsymbol{w}_j||_2}{\sqrt{s_j}a^2\xi_1},$$

which implies that $P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1}||\boldsymbol{w}_j||_1 \leqslant \frac{16\eta}{a^2\xi_1}\Big) \geqslant P(\mathcal{E}) \geqslant 1-\exp(-c_1\log p)$, and $P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1/2}||\boldsymbol{w}_j||_2 \leqslant \frac{8\eta}{a^2\xi_1}\Big) \geqslant 1-\exp(-c_1\log p)$. Since $||\boldsymbol{w}_j||_1 = ||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{d}_{0j}||_1$ and $||\boldsymbol{w}_j||_2 = ||\boldsymbol{d}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{d}_{0j}||_2$, (1) is proved.

(2) Note that $\tau_{0j}^2 = \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2 x_{i,j}\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\big\}$. We have

$$\begin{aligned} \max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}},\eta)| &\leqslant \max_{1\leqslant j\leqslant p} s_j^{-1/2}\Big|\frac{1}{n}\sum_{i=1}^n\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2 - [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\}x_{i,j}\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\Big| \\ &\quad + \max_{1\leqslant j\leqslant p} s_j^{-1/2}\Big|\frac{1}{n}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 x_{i,j}\boldsymbol{x}_i^T\boldsymbol{w}_j\Big| \\ &\quad + \max_{1\leqslant j\leqslant p} s_j^{-1/2}\Big|\frac{1}{n}\sum_{i=1}^n [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2 x_{i,j}\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j} - \tau_{0j}^2\Big| \\ &= \max_{1\leqslant j\leqslant p}|I_{j1}| + \max_{1\leqslant j\leqslant p}|I_{j2}| + \max_{1\leqslant j\leqslant p}|I_{j3}|, \end{aligned}$$

where I_{j1}, I_{j2}, I_{j3} are defined clearly in the context. Lemma B.20 implies that

$$P(\max_{1\leqslant j\leqslant p}|I_{j1}| \geqslant c_0\eta) \leqslant P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big\}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty \geqslant c_0\eta\Big) \leqslant \exp(-c_1\log p).$$

Lemma B.5 and assumption (B5) imply that there are universal positive constants c_0

and c_1 such that

$$\begin{aligned}\max_{1\leqslant j\leqslant p}|I_{j2}| &\leqslant \max_{1\leqslant i\leqslant n}[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 * \max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n s_j^{-1/2}|x_{i,j}\boldsymbol{x}_i^T\boldsymbol{w}_j| \\ &\leqslant c_0\sqrt{\max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n x_{i,j}^2}\sqrt{\max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n s_j^{-1}(\boldsymbol{x}_i^T\boldsymbol{w}_j)^2} \\ &\leqslant c_0\sqrt{\Big|\Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty}\sqrt{\max_{1\leqslant j\leqslant p}s_j^{-1}\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\boldsymbol{w}_j},\end{aligned}$$

with probability at least $1-\exp[-c_1\log(p\vee n)]$. Lemma B.13 implies that for universal positive constants c_0 and c_1,

$$\begin{aligned}P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T-\mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T)\Big|\Big|_\infty\geqslant c_0\sigma_x^2\sqrt{\frac{\log p}{n}}\Big)\leqslant\exp(-c_1\log p),\\ P\Big(\Big|\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T-\mathrm{E}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\boldsymbol{w}_j\Big|\geqslant c_0\sigma_x^2||\boldsymbol{w}_j||_2^2\sqrt{\frac{\log p}{n}}\Big)\leqslant\exp(-c_1\log p).\end{aligned}$$

Let ξ_p be the largest eigenvalue of $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)$. The results in (1) indicate that

$$\max_{1\leqslant j\leqslant p}s_j^{-1}\boldsymbol{w}_j^T\Big(\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\boldsymbol{w}_j\leqslant 2\max_{1\leqslant j\leqslant p}s_j^{-1}\boldsymbol{w}_j^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{w}_j\leqslant 2\xi_p\max_{1\leqslant j\leqslant p}s_j^{-1}||\boldsymbol{w}_j||_2^2\leqslant c_0\xi_p\eta^2$$

with probability at least $1-\exp(-c_1\log p)$. It follows that $\max_{1\leqslant j\leqslant p}|I_{j2}|\leqslant c_0\eta$ with probability at least $1-\exp(-c_1\log p)$ for universal positive constants c_0 and c_1.

To bound $\max_{1\leqslant j\leqslant p}|I_{3j}|$, as the proof of (1), we observe that $(2A_i-1)G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)x_{i,j}$ and $(2A_i-1)G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}$ are both sub-Gaussian with variance proxy no larger than $C=c_1b^2\sigma_x^2$. It is followed by the result

$$P(\max_{1\leqslant j\leqslant p}|I_{j3}|\geqslant c_0\eta)\leqslant\sum_{j=1}^p P\Big(\Big|\frac{1}{n}\sum_{i=1}^n[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2x_{i,j}\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}-\tau_{0j}^2\Big|\geqslant c_0\eta\sqrt{s_j}\Big)\leqslant p\exp(-cn\eta^2).$$

Since $\log p = o(n\eta^2)$, we can derive that

$$P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}},\eta)| \leqslant c_0\eta\Big) \leqslant \exp(-c_1\log p).$$

The first result is concluded. Lemma B.11 then implies

$$\begin{aligned} P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^{-2} \;-\; \tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta)| \geqslant c_0\eta\Big) &= P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1/2}\tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta)|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}},\eta)| \geqslant c_0\eta\Big) \\ &\leqslant P\Big(\max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_{0j}^2 - \tau_j^2(\widehat{\boldsymbol{\beta}},\eta)| \leqslant c_0\eta\Big) + P\Big(\max_{1\leqslant j\leqslant p} \tau_{0j}^{-2} * \tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta) \leqslant 1\Big) \\ &\leqslant \exp(-c_1\log p). \end{aligned}$$

(3) Observe that

$$\begin{aligned} &\max_{1\leqslant j\leqslant p} s_j^{-1}||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{\theta}_j||_1 = \max_{1\leqslant j\leqslant p} s_j^{-1}||\tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta)\boldsymbol{\phi}_j(\widehat{\boldsymbol{\beta}},\eta) - \tau_{0j}^{-2}\boldsymbol{\phi}_{0j}||_1 \\ \leqslant\; &\max_{1\leqslant j\leqslant p} \tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta)s_j^{-1}||\boldsymbol{w}_j||_1 + \max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta) - \tau_{0j}^{-2}| * s_j^{-1/2}||\boldsymbol{\phi}_{0j}||_1 \\ \leqslant\; &\max_{1\leqslant j\leqslant p} \tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta) * \max_{1\leqslant j\leqslant p} s_j^{-1}||\boldsymbol{w}_j||_1 + \max_{1\leqslant j\leqslant p} s_j^{-1/2}|\tau_j^{-2}(\widehat{\boldsymbol{\beta}},\eta) - \tau_{0j}^{-2}| * \max_{1\leqslant j\leqslant p} s_j^{-1/2}||\boldsymbol{\phi}_{0j}||_1. \end{aligned}$$

Conditional on the event that results in (1) and (2) all hold, then Lemma B.11 implies

$$\max_{1\leqslant j\leqslant p} s_j^{-1}||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{\theta}_j||_1 \leqslant c_1\eta * (M_2/a^2 + \widetilde{s}^{1/2}\eta) + c_1\eta * b^2M_1M_2/a^2 \leqslant c_0\eta.$$

Similar proofs can be applied for $\max_{1\leqslant j\leqslant p} s_j^{-1}||\boldsymbol{\theta}_j(\widehat{\boldsymbol{\beta}},\eta) - \boldsymbol{\theta}_j||_2$.

(4) Let $\mathbb{K}(s) = \{\boldsymbol{v}\in\mathbb{R}^p : ||\boldsymbol{v}||_0 \leqslant s, ||\boldsymbol{v}||_2 = 1\}$. By taking $s = \frac{n}{2\log p}$ and $t = \frac{\xi_1}{54}$, Lemma B.14 implies that

$$P\Big(\sup_{\boldsymbol{v}\in\mathbb{K}(2s)} \Big|\boldsymbol{v}^T\Big[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{x}_i^T)\Big]\boldsymbol{v}\Big| \geqslant \frac{\xi_1}{54}\Big) \leqslant 2\exp\Big(-cn\min\Big\{\frac{\xi_1}{54}, 1\Big\}\Big).$$

Then Lemma 13 in Loh and Wainwright (2012) implies that for any $\boldsymbol{v}\in\mathbb{R}^p$, we have that $\boldsymbol{v}^T\big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\big)\boldsymbol{v} \leqslant \frac{3\xi_p}{2}||\boldsymbol{v}||_2^2 + \frac{\xi_1\log p}{n}||\boldsymbol{v}||_1^2$.

Results in (3) imply that $\max_{1\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_2^2 \leqslant 2\max_{1\leqslant j\leqslant p}(||\boldsymbol{\theta}_j||_2^2 + ||\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j||_2^2) \leqslant 2(M_2^2/a^4 + c_0\eta^2\widetilde{s})$ and $\max_{1\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j||_1^2 \leqslant 2\max_{1\leqslant j\leqslant p}(||\boldsymbol{\theta}_j||_1^2 + ||\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j||_1^2) \leqslant 2(\widetilde{s}M_2^2/a^4 + c_0\widetilde{s}^2\eta^2)$, with probability at least $1-\exp(-c_1\log p)$. Hence we can derive that

$$\begin{aligned}\max_{1\leqslant j\leqslant p}\widehat{\boldsymbol{\theta}}_j^T\Big(\frac{1}{n}\sum_{i=1}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\widehat{\boldsymbol{\theta}}_j &\leqslant \frac{3\xi_p}{2}||\widehat{\boldsymbol{\theta}}_j||_2^2 + \frac{\xi_1\log p}{n}||\widehat{\boldsymbol{\theta}}_j||_1^2 \\ &\leqslant 3\xi_p(M_2^2/a^4 + c_0h^2\widetilde{s}) + \frac{2\xi_1\log p}{n}(\widetilde{s}M_2^2/a^4 + c_0\widetilde{s}^2h^2) \\ &\leqslant 4\xi_pM_2^2/a^4,\end{aligned}$$

since $h^2\widetilde{s} = o(1)$, $n^{-1}\widetilde{s}\log p = o(h^5\widetilde{s}) = o(1)$ and $n^{-1}\widetilde{s}^2h^2\log p = o(h^7\widetilde{s}^2) = o(1)$. □

Proof of Theorem 3.2 Observing that

$$\begin{aligned}\boldsymbol{S}_n(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{S}_n(\boldsymbol{\beta}_0) - \frac{1}{n}\sum_{i=1}^n\big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i \\ &\quad -\frac{1}{n}\sum_{i=1}^n\big[\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\boldsymbol{x}_i \\ &= \boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{J}_1 - \boldsymbol{J}_2,\end{aligned}$$

where the definition of $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$ is clear from the context, we have

$$\begin{aligned}\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \sqrt{n}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \sqrt{n}\widehat{\boldsymbol{\Theta}}^T\big[\boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{J}_1 - \boldsymbol{J}_2\big] \\ &= \sqrt{n}\big\{\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \widehat{\boldsymbol{\Theta}}^T\big[\boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{J}_1 - \boldsymbol{J}_2\big] - \boldsymbol{\Theta}^T\boldsymbol{D}_{n1}\big\} + \sqrt{n}\boldsymbol{\Theta}^T\boldsymbol{D}_{n1} \\ &= \boldsymbol{\Delta} + \boldsymbol{W},\end{aligned}$$

where $\boldsymbol{D}_{n1} = n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_iG^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i$. It is easy to see that $\sqrt{n}\boldsymbol{D}_{n1} \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Lambda})$. Hence to prove the theorem, it suffices to show that $P(||\boldsymbol{\Delta}||_\infty \geqslant c_0\Delta_{n,p}) \leqslant \exp[-c_1\log(p\wedge n)]$.

We will first show for $\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_2||_\infty$. Note that $\boldsymbol{J}_2 = \boldsymbol{J}_{21} + \boldsymbol{J}_{22} - \boldsymbol{J}_{23}$, where

$$\begin{aligned}
\boldsymbol{J}_{21} &= \frac{1}{n}\sum_{i=1}^n \big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\boldsymbol{x}_i,\\
\boldsymbol{J}_{22} &= \frac{1}{n}\sum_{i=1}^n \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\widetilde{\epsilon}_i\boldsymbol{x}_i,\\
\boldsymbol{J}_{23} &= \frac{1}{n}\sum_{i=1}^n \big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\widetilde{\epsilon}_i\boldsymbol{x}_i.
\end{aligned}$$

First, note that Lemma 3.2-(4) implies that with probability at least $1-\exp[-c_1\log(p\wedge n)]$, via Cauchy-Schwartz inequality,

$$\max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j| \leqslant \max_{1\leqslant j\leqslant p}\sqrt{\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j|^2} \leqslant 2\sqrt{\xi_p}M_2/a^2.$$

Similarly, we can also prove that $\max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T(\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}_j)| \leqslant \max_{1\leqslant j\leqslant p}\sqrt{\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T(\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}_j)|^2} \leqslant 2\eta\sqrt{\widetilde{s}\xi_p}$.

Lemma B.10 and Lemma 3.2 together imply that

$$\begin{aligned}
||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_{21}||_\infty &\leqslant \max_{1\leqslant i\leqslant n}|\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)| * \max_{1\leqslant i\leqslant n}|\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)| * \max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j|\\
&= O_p\big(sh^2\sqrt{\log(p\vee n)}\big) * O_p(h) * O_p(1) = O_p\big(sh^3\sqrt{\log(p\vee n)}\big).
\end{aligned}$$

For $\boldsymbol{J}_{22}$, Lemma B.7, Lemma B.11 and Lemma 3.2 together imply that $\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_{22}||_\infty \leqslant \sqrt{n}||\boldsymbol{\Theta}^T\boldsymbol{J}_{22}||_\infty + \sqrt{n}||\boldsymbol{J}_{22}||_\infty * \max_{1\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j||_1 = O_p(\gamma_1)$, with $\gamma_1 = \sqrt{h\log(p\vee n)} + h[\log(p\vee n)]^{3/2}$. Similarly, $\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_{23}||_\infty = O_p(\gamma_1)$. Note that $\sqrt{h\log(p\vee n)} = o(h^3\sqrt{n\log(p\vee n)})$. Hence we can conclude that $P(\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_2||_\infty \geqslant c_0\Delta_{n,p}) \leqslant P\Big(\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_2||_\infty \geqslant c_0\big[\gamma_1 + sh^3\sqrt{n\log(p\vee n)}\big]\Big) \leqslant \exp[-c_1\log(p\vee n)]$.

We next show for $\sqrt{n}||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + \widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1||_\infty$. Note that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \;+\; & \widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1 = \big\{\boldsymbol{I} - \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\big\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
& + \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n \big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big]\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i \\
= \; & \big\{\boldsymbol{I} - \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\big\}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
& + \boldsymbol{\Theta}^T n^{-1}\sum_{i=1}^n \big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i \\
& + (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T n^{-1}\sum_{i=1}^n \big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big]G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i \\
+ \; & \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n \big[\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big]\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]\boldsymbol{x}_i \\
= \; & \boldsymbol{I}_1 + \boldsymbol{I}_2 + \boldsymbol{I}_3 + \boldsymbol{I}_4,
\end{aligned}$$

where the definition of $\boldsymbol{I}_i = (I_{i1}, \ldots, I_{ip})^T$ $(i = 1, \ldots, 4)$ is clear from the context. The definition of $\widehat{\boldsymbol{\theta}}_j$, Lemma B.11 and Lemma 3.2 imply that

$$\Big|\Big|n^{-1}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \leqslant \eta\widehat{\tau}_j^{-2} = \eta[\tau_j^{-2} + O_p(\sqrt{s_j}\eta)] = O_p(\eta),$$

uniformly in j. Note that $\Big|1 - n^{-1}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j x_{i,j}\Big| = |1 - \widehat{\tau}_j^{-2} * \widehat{\tau}_j^2| = 0$, hence we can conclude that $\Big|\Big|\boldsymbol{I} - \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n [\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty = O_p(\eta)$, and therefore $||\boldsymbol{I}_1||_\infty = O_p(\eta) * ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 = O_p(s\lambda\eta) = O_p\big(sh^3\big)$.

Lemma B.8 implies that $\sqrt{n}||\boldsymbol{I}_2||_\infty \leqslant O_p(h^{3/4})$, and $\sqrt{n}||\boldsymbol{I}_3||_\infty \leqslant \max_{1\leqslant j\leqslant p}||\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j||_1 * O_p(h^{3/4}) = O_p(h^{3/4})$. For $\boldsymbol{I}_4$, first note that existing lemmas and assumptions to-

gether imply that

$$\begin{aligned}
& \big|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - \widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{x}_i^T(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})\big| \\
\leqslant\ & \big|\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big| + \big|\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) - G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big| + O_p(1) * \max_{1\leqslant i\leqslant n}|\boldsymbol{x}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| \\
=\ & O_p\big(sh^2\sqrt{\log(p\vee n)}\big).
\end{aligned}$$

Hence we can conclude that

$$||\boldsymbol{I}_4||_\infty \leqslant O_p\big(sh^2\sqrt{\log(p\vee n)}\big) * O_p(h) * \max_{1\leqslant j\leqslant p}\frac{1}{n}\sum_{i=1}^n|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j| = O_p\big(sh^3\sqrt{\log(p\vee n)}\big).$$

All these results together indicate that $P(\sqrt{n}||\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0+\widehat{\boldsymbol{\Theta}}^T\boldsymbol{J}_1||_\infty \geqslant c_0\Delta_{n,p}) \leqslant \exp[-c_1\log(p\wedge n)]$.

We next examine $-\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{\Theta}^T\boldsymbol{D}_{n1}$. Note that

$$\begin{aligned}
& -\widehat{\boldsymbol{\Theta}}^T\boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{\Theta}^T\boldsymbol{D}_{n1} \\
=\ & \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n\big\{\widetilde{\epsilon}_i + f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - \widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\widehat{G}^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)\boldsymbol{x}_i - \boldsymbol{\Theta}^T n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i \\
=\ & (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^T n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i + \widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i\big\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big\}\boldsymbol{x}_i \\
& -\widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n\big\{\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)\boldsymbol{x}_i \\
& -\widehat{\boldsymbol{\Theta}}^T n^{-1}\sum_{i=1}^n\big\{\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\big\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big\}\boldsymbol{x}_i \\
=\ & \sum_{i=2}^5\boldsymbol{D}_{ni},
\end{aligned}$$

where $\boldsymbol{D}_{ni}$'s are defined clearly from the context. Hence it suffices to show that $\sqrt{n}||\boldsymbol{D}_{ni}||_\infty = O_p(\Delta_{n,p})$ for $i = 2,\cdots,5$.

For $\boldsymbol{D}_{n2}$, note that given $(\epsilon_i,\boldsymbol{x}_i)$, for any j, $n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$ is sub-

Gaussian with variance proxy $b^2 n^{-1} \sum_{i=1}^n \big(\epsilon_i + g(\boldsymbol{x}_i)\big)^2 [\boldsymbol{x}_i^T(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)]^2 \xrightarrow{p} c\widetilde{s}h^2$, for constant $c > 0$. Hence $\sqrt{n}||\boldsymbol{D}_{n2}||_\infty \leqslant ch\sqrt{\widetilde{s}\log p}$ with probability at least $1 - \exp(-c_1 \log p)$. Similarly, $\sqrt{n}||\boldsymbol{D}_{n3}||_\infty \leqslant ch\sqrt{\log p}$ with the same probability bound.

Lemma B.9 indicates that $\sqrt{n}||\boldsymbol{D}_{n4}||_\infty = O_p(h^{3/4})$. For $\boldsymbol{D}_{n5}$, Lemma B.10 and Lemma 3.2 together imply that

$$||\boldsymbol{D}_{n5}||_\infty \leqslant O_p\big(sh^2\sqrt{\log(p \vee n)}\big) * O_p(h) * O_p(1) = O_p\big(sh^3\sqrt{\log(p \vee n)}\big).$$

All these results imply that $P(\sqrt{n}||\widehat{\boldsymbol{\Theta}}^T \boldsymbol{S}_n(\boldsymbol{\beta}_0) - \boldsymbol{\Theta}^T \boldsymbol{D}_{n1}||_\infty \geqslant c_0 \Delta_{n,p}) \leqslant \exp[-c_1 \log(p \vee n)]$, which finishes the proof. □

Proof of Corollary 3.1 Let $\sigma_j^2 = \mathrm{E}\big\{\big[\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j\big]^2\big\}$, which is the j^{th} diagonal entry of $\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}$. Lemma B.11 implies that

$$\sigma_\epsilon^2 a^2 \xi_1/(b^4 M_1^2) \leqslant \sigma_\epsilon^2 a^2 \xi_1 ||\boldsymbol{\theta}_j||_2^2 \leqslant \sigma_j^2 \leqslant (\sigma_\epsilon^2 + M_g^2) b^2 \xi_p ||\boldsymbol{\theta}_j||_2^2 \leqslant (\sigma_\epsilon^2 + M_g^2) b^2 \xi_p M_2^2/a^4,$$

which suggests that $\max_{1\leqslant j\leqslant p} \sigma_j^{-1} \leqslant b^2 M_1/(a\sigma_\epsilon\sqrt{\xi_1})$. Since Lemma B.12 show that $\max_{1\leqslant j\leqslant p} \big|\widehat{\Sigma}_{j,j} - \sigma_j^2\big| = o_p(1)$, we can derive that $\max_{1\leqslant j\leqslant p} \big|\widehat{\Sigma}_{j,j}^{-1/2} - \sigma_j^{-1}\big| = o_p(1)$.

Note that Theorem 3.2 already show that $\max_{1\leqslant j\leqslant p} \Big|\widetilde{\beta}_j - \beta_{0j} - \frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j\Big| = o_p(n^{-1/2})$. Hence we have

$$\begin{aligned}
&\max_{1\leqslant j\leqslant p} \Big|\widehat{\Sigma}_{j,j}^{-1/2}(\widetilde{\beta}_j - \beta_{0j}) - \sigma_j^{-1}\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j\Big| \\
\leqslant\ & \max_{1\leqslant j\leqslant p} \sigma_j^{-1} * \max_{1\leqslant j\leqslant p} \Big|\widetilde{\beta}_j - \beta_{0j} - \frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j\Big| + \max_{1\leqslant j\leqslant p} \big|\widehat{\Sigma}_{j,j}^{-1/2} - \sigma_j^{-1}\big| * \max_{1\leqslant j\leqslant p} |\widetilde{\beta}_j - \beta_{0j}| \\
=\ & o_p(n^{-1/2}) * O(1) + o_p(1) * O_p(n^{-1/2}) = o_p(n^{-1/2}).
\end{aligned}$$

Applying the Berry-Esseen bound for CLT, there is a universal constant $c_0 > 0$ such

that

$$\begin{aligned}
&\max_{1\leqslant j\leqslant p}\sup_{\alpha\in(0,1)}\Big|P\Big(\sqrt{n}|\sigma_j^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j|\leqslant\Phi^{-1}(1-\alpha/2)\Big)-(1-\alpha)\Big|\\
\leqslant\ &\frac{c_0}{\sqrt{n}}\mathrm{E}\big(|\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j|^3\big)\leqslant\frac{c_0 b^3}{\sqrt{n}}\mathrm{E}\big(|\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j|^3\big).
\end{aligned}$$

Recall that $\widetilde{\epsilon}_i = 2(2A_i-1)\big(\epsilon_i + g(\boldsymbol{x}_i)\big)$, where ϵ_i is sub-Gaussian independent of $\boldsymbol{x}_i$, and $P\big(g(\boldsymbol{x}_i)\leqslant M_g\big)=1$, then we can conclude that above probability is bounded by $\frac{c_1}{\sqrt{n}}\mathrm{E}\big(|\boldsymbol{x}_i^T\boldsymbol{\theta}_j|^3\big)$, where c_1 does not dependent on n, p and $\boldsymbol{\beta}_0$. Note that $\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy $\sigma_x^2||\boldsymbol{\theta}_j||_2^2$. Lemma B.11 implies that $\mathrm{E}\big(|\boldsymbol{x}_i^T\boldsymbol{\theta}_j|^3\big)\leqslant c_2||\boldsymbol{\theta}_j||_2^3\leqslant c_2(M_2/a^2)^3$, where $c_2>0$ is does not dependent on n, p and $\boldsymbol{\beta}_0$. Then for a universal constant $c>0$,

$$\max_{1\leqslant j\leqslant p}\sup_{\alpha\in(0,1)}\Big|P\Big(\sqrt{n}|\sigma_j^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j|\leqslant\Phi^{-1}(1-\alpha/2)\Big)-(1-\alpha)\Big|\leqslant\frac{c}{\sqrt{n}}=o(1).$$

Combining all the results above, we conclude the corollary. □

Proof of Theorem 3.3 Let $\widetilde{\delta}_j = n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j$, and $\boldsymbol{\xi}=(\xi_1,\cdots,\xi_p)^T$ be a multivariate mean zero Gaussian with covariance matrix $\boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}$.

Since $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)$ and $\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ are both sub-Gaussian, Comment 2.2 in Chakraborty et al. (2014) implies that $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ satisfies condition (E.1) with $B_n=C_1$ for some universal constant C_1, which does not depend on n, p and $\boldsymbol{\beta}_0$. The order of h implies that $\log p = o\big((\sqrt{n}h^3)^{-1}\big)$, and $(nh^5)^{-1}=o(1)$. Hence we can derive that $\frac{[\log(pn)]^7}{n}=O\Big(\frac{[\log p]^7}{n}\Big)=o\big((n^{9/2}h^{21})^{-1}\big)=o(n^{-0.3})$. Note that $\max_{1\leqslant j\leqslant p}|\widetilde{\delta}_j|=\max_{1\leqslant j\leqslant p}\{\widetilde{\delta}_j,-\widetilde{\delta}_j\}$, hence Corollary 2.1 in Chernozhukov et al. (2013) indicates that $\sup_{t\in\mathbb{R}}\big|P(\sqrt{n}\max_{1\leqslant j\leqslant p}|\widetilde{\delta}_j|\leqslant t)-P(\max_{1\leqslant j\leqslant p}|\xi_j|\leqslant t)\big|\leqslant\exp(-c_1\log n)$, for universal constant $c_1>0$.

Define the event $\mathbb{T}_n(\mathcal{G})=\{\sqrt{n}\big|\max_{j\in\mathcal{G}}|\widetilde{\beta}_j-\beta_{0j}|-\max_{j\in\mathcal{G}}|\widetilde{\delta}_j|\big|>\Delta_{n,p}\}$. Note that

Theorem 3.2 implies for universal constant $c_1 > 0$,

$$P\big(\mathbb{T}_n(\mathcal{G})\big) \leqslant P\Big(\sqrt{n}\big|\max_{j\in\mathcal{G}}|(\widetilde{\beta}_j - \beta_{0j}) - \widetilde{\delta}_j|\big| > \Delta_{n,p}\Big) \leqslant P\Big(||\boldsymbol{\Delta}||_\infty \geqslant \Delta_{n,p}\Big) \leqslant \exp(-c_1 \log p).$$

Applying Corollary 16 in Wasserman (2014), we have that for universal constant $C > 0$,

$$\begin{aligned} P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j - \beta_{0j}| \leqslant c^*_{1-\alpha}\Big) &\leqslant P\Big(\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| \leqslant c^*_{1-\alpha} + \Delta_{n,p}\Big) + P\big(\mathbb{T}_n(\mathcal{G})\big) \\ &\leqslant P\Big(\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| \leqslant c^*_{1-\alpha}\Big) + C\Delta_{n,p}\sqrt{1 \vee \log(p/\Delta_{n,p})} + \exp(-c_1 \log p), \end{aligned}$$

uniformly over $\alpha \in (0,1)$. Note that $\Delta_{n,p}\sqrt{\log p} = o(1)$ implies that $\Delta_{n,p}\sqrt{1 \vee \log(p/\Delta_{n,p})} = o(1)$. Hence it suffices to show $P\Big(\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| \leqslant c^*_{1-\alpha}\Big) \leqslant 1 - \alpha + o(1)$ uniformly over $\alpha \in (0,1)$.

Note that $\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| = \max_{j\in\mathcal{G}}\{\widetilde{\delta}_j, -\widetilde{\delta}_j\}$, and $\max_{j\in\mathcal{G}}|\delta^*_j| = \max_{j\in\mathcal{G}}\{\delta^*_j, -\delta^*_j\}$. Observe that conditional on w, $(\delta^*_1, \cdots, \delta^*_p, -\delta^*_1, \cdots, -\delta^*_p)^T$ is multivariate mean zero Gaussian with covariance matrix $\begin{pmatrix} \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) & -\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \\ -\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) & \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) \end{pmatrix}$. Then let $\Delta_0 = ||\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\Theta}^T\boldsymbol{\Lambda}\boldsymbol{\Theta}||_\infty$, Gaussian comparison inequality suggests that

$$\begin{aligned} P\Big(\max_{j\in\mathcal{G}}|\widetilde{\delta}_j| \leqslant c^*_{1-\alpha}\Big) &\leqslant P\Big(\max_{j\in\mathcal{G}}|\delta^*_j| \leqslant c^*_{1-\alpha}\big|w\Big) + C\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3} \\ &= 1 - \alpha + C\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3}, \end{aligned}$$

uniformly over $\alpha \in (0,1)$. Proof of Lemma B.12 implies that $\Delta_0 \leqslant c_0\widetilde{s}^{1/2}h$ with probability at least $1 - \exp(-c_1 \log p)$, for universal positive constants c_0 and c_1. Hence, we can derive $\Delta_0 \log^2 p = o_p(1)$. It implies that $\Delta_0^{1/3}\big[1 \vee \log(p/\Delta_0)\big]^{2/3} = o(1)$ for all sufficiently large n. We obtain

$$\sup_{\alpha\in(0,1)}\Big[P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j - \beta_{0j}| \leqslant c^*_{1-\alpha}\Big) - (1-\alpha)\Big] \leqslant \exp[-c_1 \log(p \wedge n)].$$

Similarly, we can derive that $\sup_{\alpha\in(0,1)}\Big[(1-\alpha)-P\Big(\sqrt{n}\max_{j\in\mathcal{G}}|\widetilde{\beta}_j-\beta_{0j}|\leqslant c^*_{1-\alpha}\Big)\Big]\leqslant 1-\exp(-c_1\log p)$. Note that all the universal constants do not depend on n, p and $\boldsymbol{\beta}_0$. It concludes the theorem. □

B.5 Proofs of Technical Lemmas in Appendix B.2

Proof of Lemma B.1 By the model setup, we have:

$$\begin{aligned}
\boldsymbol{S}_n(\boldsymbol{\beta}_0) &= -n^{-1}\sum_{i=1}^n\big\{\widetilde{\epsilon}_i+G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i\\
&= -n^{-1}\sum_{i=1}^n G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\widetilde{\epsilon}_i\boldsymbol{x}_i-n^{-1}\sum_{i=1}^n\big\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\widetilde{\epsilon}_i\boldsymbol{x}_i\\
&\quad -n^{-1}\sum_{i=1}^n\big\{G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i\\
&\quad -n^{-1}\sum_{i=1}^n\big\{G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\big\{\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big\}\boldsymbol{x}_i\\
&= \sum_{j=1}^4\boldsymbol{I}_{nj},
\end{aligned}$$

where $\boldsymbol{I}_{nj}$'s are defined clearly from the context. Since $G^{(1)}(t,\boldsymbol{\beta})$ is bounded, and $\widetilde{\epsilon}_i$ is sub-Gaussian by Lemma B.2, then $G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\widetilde{\epsilon}_i$ is also sub-Gaussian. Lemma B.13 implies that there exist positive constants c_0, c_1 and c_2 such that

$$P\Big(||\boldsymbol{I}_{n1}||_\infty\geqslant c_0\sqrt{\frac{\log p}{n}}\Big)=P\Big(\big|\big|n^{-1}\sum_{i=1}^n\widetilde{\epsilon}_i\boldsymbol{x}_i\big|\big|_\infty>c_0\sqrt{\frac{\log p}{n}}\Big)\leqslant c_1\exp(-c_2\log p).$$

Similarly, Lemma B.5 indicates with probability at least $1-\exp[-c_0 s\log(p\vee n)]$, $\widetilde{\epsilon}_i\big[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big]$ is sub-Gaussian. Hence we can derive that $P\Big(||\boldsymbol{I}_{n2}||_\infty\geqslant d_0h\sqrt{\frac{\log p}{n}}\Big)\leqslant\exp(-c\log p)$.

Lemma B.9 implies that $P\big(\sqrt{n}||\boldsymbol{I}_{n3}||_\infty \geqslant h^{3/4}\big) \leqslant \exp[-c_1\log(p\vee n)]$. Hence $||\boldsymbol{I}_{n3}||_\infty \leqslant d_0 n^{-1/2}$ with probability at least $1-\exp(-c\log p)$.

For $\boldsymbol{I}_{n4}$, note that $\boldsymbol{x}_i$ is sub-Gaussian. Lemma B.13 indicates that for universal constant $c_1>0$,

$$\begin{aligned} P\Big(\frac{1}{n}\sum_{i=1}^n||\boldsymbol{x}_i\boldsymbol{x}_i^T||_\infty \geqslant ||\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)||_\infty + \sigma_x^2\sqrt{\frac{\log p}{n}}\Big) &\leqslant P\Big(||\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T - \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)||_\infty \geqslant \sigma_x^2\sqrt{\frac{\log p}{n}}\Big) \\ &\leqslant \exp(-c_1\log p). \end{aligned}$$

Lemma B.4 and Lemma B.5 together indicate $||\boldsymbol{I}_{n4}||_\infty \leqslant ch^3 n^{-1}\sum_{i=1}^n||\boldsymbol{x}_i||_\infty \leqslant ch^3\sqrt{\frac{1}{n}\sum_{i=1}^n||\boldsymbol{x}_i\boldsymbol{x}_i^T||_\infty} \leqslant c_0 n^{-1/2}$ with probability at least $1-\exp(-c_1\log p)$. Combining all the previous results, we conclude the lemma. □

Proof of Lemma B.2 For any unit vector $\boldsymbol{v}\in\mathbb{R}^p$ and $c\in\mathbb{R}$, Jensen's inequality and the sub-Gaussian property for $\boldsymbol{x}$ imply that

$$\begin{aligned} \mathrm{E}\big\{\exp[s\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)]\big\} &= \mathrm{E}\big\{\exp[\mathrm{E}(s\boldsymbol{x}^T\boldsymbol{v}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)]\big\} \leqslant \mathrm{E}\big\{\mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})\big|\boldsymbol{x}^T\boldsymbol{\beta}_0]\big\} \\ &= b\mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})] \leqslant \exp\big(\frac{s^2\sigma^2}{2}\big). \end{aligned}$$

To show for $\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)$, we apply Hölder's inequality, which indicates that for any $\frac{1}{p}+\frac{1}{q}=1$,

$$\begin{aligned} \mathrm{E}\Big\{\exp\big\{s\boldsymbol{v}^T[\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)]\big\}\Big\} &\leqslant \big[\mathrm{E}\exp(sp\boldsymbol{x}^T\boldsymbol{v})\big]^{1/p}\Big\{\mathrm{E}\exp\big[sq\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big]\Big\}^{1/q} \\ &\leqslant \mathrm{E}\quad \exp(sp\boldsymbol{x}^T\boldsymbol{v}) * \mathrm{E}\exp\big[sq\boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)\big] \leqslant \exp\big[\frac{s^2\sigma^2(p^2+q^2)}{2}\big]. \end{aligned}$$

Let $p=q=2$, then we have $\mathrm{E}\Big\{\exp\big\{s\boldsymbol{v}^T[\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}\big|\boldsymbol{x}^T\boldsymbol{\beta}_0)]\big\}\Big\} \leqslant \exp(2s^2\sigma^2)$.

For $\widetilde{\epsilon} = 2(2A-1)\big(\epsilon+g(\boldsymbol{x})\big)$, note that ϵ and A are independent. Easy to show that $2(2A-1)\epsilon$ is sub-Gaussian, since $\mathrm{E}\big\{\exp[2s(2A-1)\epsilon]\big\} = \frac{1}{2}\big(\mathrm{E}e^{2s\epsilon}+\mathrm{E}e^{-2s\epsilon}\big) \leqslant$

$\exp(2s^2\sigma_\epsilon^2)$. Since $g(\cdot)$ is bounded by M_g almost everywhere, Exercise 2.4 in Wainwright (2015) implies that $2(2A-1)g(\boldsymbol{x})$ is sub-Gaussian with variance proxy at most $4M_g^2$. Then similar as the previous step, we can conclude that $\widetilde{\epsilon}$ is sub-Gaussian. □

Proof of Lemma B.3 Note that $P(|g(\boldsymbol{x}_i)| \leqslant M_g) = 1$, and $\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy $||\boldsymbol{\theta}_j||_2^2\sigma_x^2$, where $||\boldsymbol{\theta}_j||_2 \leqslant M_2/a^2$ uniformly in j by Lemma B.11. Then for sufficiently large n, probability bounds for $\mathcal{G}_n$ and $\mathcal{K}_n$ can be easily derived from the property of sub-Gaussian, in Definition 1. For $\mathcal{H}_n$, Lemma B.2 implies that $\widetilde{\epsilon}_i$ is also sub-Gaussian. Note that $\mathrm{E}(\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j) = 0$. Lemma B.13 implies that there exists universal constant $c > 0$ such that

$$\begin{aligned} & P\Big(\max_{1\leqslant j\leqslant p, 1\leqslant i\leqslant n} |\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j| \geqslant \sigma_x(\sigma_\epsilon + M_g)\log(p\vee n)\Big) \\ \leqslant & \sum_{i=1}^n\sum_{j=1}^p P\Big(|\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j| \geqslant \sigma_x(\sigma_\epsilon + M_g)\log(p\vee n)\Big) \\ \leqslant & np\exp[-c_1\log(p\vee n)] = \exp[-c\log(p\vee n)]. \end{aligned}$$

□

Proof of Lemma B.4 Note that

$$\begin{aligned} \widehat{G}(t|\boldsymbol{\beta}) - G(t|\boldsymbol{\beta}) &= \sum_{i=1}^n W_{ni}(t|\boldsymbol{\beta})\big(\widetilde{Y}_i - G(t|\boldsymbol{\beta})\big) \\ &= \frac{n^{-1}\sum_{i=1}^n K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\big[\widetilde{Y}_i - G(t|\boldsymbol{\beta})\big]}{n^{-1}\sum_{i=1}^n K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})} \\ &= \frac{A_{n1}(t|\boldsymbol{\beta}) + A_{n2}(t|\boldsymbol{\beta})}{A_{n3}(t|\boldsymbol{\beta})}, \end{aligned}$$

where $A_{n1}(t|\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\widetilde{\epsilon}_i$, $A_{n2}(t|\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G(t|\boldsymbol{\beta})\big]$, and $A_{n3}(t|\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})$. Then Lemma B.16–B.18 show the high

probability bounds for A_{ni}, $i = 1, 2, 3$. Assumption (B4) show that $f_{\boldsymbol{\beta}}^{-1}(t) = O(1)$ for all t and $\boldsymbol{\beta}$. Then we can derive that for universal positive constants c_0 and c_1, we can conclude

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} \big|\widehat{G}(t|\boldsymbol{\beta}) - G(t|\boldsymbol{\beta})\big| \geqslant c_0 h^2 \Big) \leqslant \exp(-c_1 n h^5).$$

□

Proof of Lemma B.5 Since we know

$$\widehat{G}(t|\boldsymbol{\beta}) - G(t|\boldsymbol{\beta}) = \frac{A_{n1}(t|\boldsymbol{\beta}) + A_{n2}(t|\boldsymbol{\beta})}{A_{n3}(t|\boldsymbol{\beta})},$$

then we have

$$\widehat{G}^{(1)}(t|\boldsymbol{\beta}) - G^{(1)}(t|\boldsymbol{\beta}) = \frac{A_{n1}^{(1)}(t|\boldsymbol{\beta}) + A_{n2}^{(1)}(t|\boldsymbol{\beta})}{A_{n3}(t|\boldsymbol{\beta})} + \frac{A_{n1}(t|\boldsymbol{\beta}) + A_{n2}(t|\boldsymbol{\beta})}{A_{n3}(t|\boldsymbol{\beta})} * \frac{A_{n3}^{(1)}(t|\boldsymbol{\beta})}{A_{n3}(t|\boldsymbol{\beta})},$$

where $A_{nj}^{(1)}(t|\boldsymbol{\beta}) = \frac{d}{dt} A_{nj}(t|\boldsymbol{\beta})$, for $j = 1, 2, 3$. Let $K_h(z) = h^{-1}K(z/h)$, and $K_h'(z) = h^{-2}K'(z/h)$. Then we have

$$\begin{aligned}
A_{n1}^{(1)}(t|\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} K_h'(t - \boldsymbol{x}_i^T\boldsymbol{\beta})\widetilde{\epsilon}_i, \\
A_{n2}^{(1)}(t|\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} K_h'(t - \boldsymbol{x}_i^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - G(t|\boldsymbol{\beta})\big] - G^{(1)}(t|\boldsymbol{\beta}) n^{-1}\sum_{i=1}^{n} K_h(t - \boldsymbol{x}_i^T\boldsymbol{\beta}) \\
&= A_{n21}^{(1)}(t|\boldsymbol{\beta}) - A_{n22}^{(1)}(t|\boldsymbol{\beta}), \\
A_{n3}^{(1)}(t|\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} K_h'(t - \boldsymbol{x}_i^T\boldsymbol{\beta}).
\end{aligned}$$

First, similarly as in the proof of Lemma B.16, we can derive that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} |A_{n1}^{(1)}(t|\boldsymbol{\beta})| \geqslant c_0 h \Big) \leqslant \exp(-c_1 n h^5).$$

Observe that

$$\begin{aligned}
\mathrm{E}K_h'(t-\boldsymbol{x}_i^T\boldsymbol{\beta}) &= h^{-2}\int K'\Big(\frac{t-y}{h}\Big)f_{\boldsymbol{\beta}}(y)dy \\
&= h^{-1}\int K'(-z)\Big[f_{\boldsymbol{\beta}}(t)+hzf_{\boldsymbol{\beta}}'(t)+\frac{h^2z^2}{2}f_{\boldsymbol{\beta}}''(\widetilde{t})\Big]dz \\
&= f_{\boldsymbol{\beta}}'(t)+\frac{h}{2}\int z^2K'(-z)f_{\boldsymbol{\beta}}''(\widetilde{t})dz.
\end{aligned}$$

Similarly as in the proof of Lemma B.18, for universal positive constants c_0 and c_1,

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n3}^{(1)}(t|\boldsymbol{\beta})-f_{\boldsymbol{\beta}}'(t)\big|\geqslant c_0h\Big)\leqslant\exp(-c_1nh^5).$$

Then the techniques in the proof of Lemma B.16 and Lemma B.17 can be applied to analyze $A_{n21}^{(1)}(t|\boldsymbol{\beta})$ and $A_{n22}^{(1)}(t|\boldsymbol{\beta})$. Observe that for some t_1 and t_2 between t and $t+hz$,

$$\begin{aligned}
&\mathrm{E}K_h'(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\big[f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(t|\boldsymbol{\beta})\big] \\
=\;&\mathrm{E}K_h'(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\big[G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(t|\boldsymbol{\beta})\big] \\
=\;&h^{-1}\int K'(-z)\big[G(t+hz|\boldsymbol{\beta})-G(t|\boldsymbol{\beta})\big]f_{\boldsymbol{\beta}}(t+hz)dz \\
=\;&-\int K'(z)\Big[zG^{(1)}(t|\boldsymbol{\beta})+\frac{hz^2}{2}G^{(2)}(t_1|\boldsymbol{\beta})\Big]\Big[f_{\boldsymbol{\beta}}(t)+hzf_{\boldsymbol{\beta}}'(t_2)\Big]dz \\
=\;&G^{(1)}(t|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(t)-\frac{hf_{\boldsymbol{\beta}}(t)}{2}\int z^2K'(z)G^{(2)}(t_1|\boldsymbol{\beta})dz \\
&-hG^{(1)}(t|\boldsymbol{\beta})\int z^2K'(z)f_{\boldsymbol{\beta}}'(t_2)dz-\frac{h^2}{2}\int z^3K'(z)G^{(2)}(t_1|\boldsymbol{\beta})f_{\boldsymbol{\beta}}'(t_2)dz.
\end{aligned}$$

Then assumption (B3)–(B5) and the proofs in Lemma B.17 and Lemma B.18 imply that for some constants c_0 and c_1,

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n21}^{(1)}(t|\boldsymbol{\beta})-G^{(1)}(t|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(t)\big|\geqslant c_0h\Big)\leqslant\exp(-c_1nh^5),$$

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} \big|A^{(1)}_{n22}(t|\boldsymbol{\beta}) - G^{(1)}(t|\boldsymbol{\beta})f_{\boldsymbol{\beta}}(t)\big| \geqslant c_0 h\Big) \leqslant \exp(-c_1 n h^5).$$

This implies that $P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} \big|A^{(1)}_{n2}(t|\boldsymbol{\beta})\big| \leqslant c_0 h\Big) \geqslant 1 - 2\exp(-c_1 n h^5)$. Assumption (B4) show that $f^{-1}_{\boldsymbol{\beta}}(t) = O(1)$ for all t and $\boldsymbol{\beta}$. Then given the high probability bounds for $A_{ni}(t|\boldsymbol{\beta})$ and $A^{(1)}_{ni}(t|\boldsymbol{\beta})$, we conclude the lemma. □

Proof of Lemma B.6 Note that

$$\begin{aligned} P\big(\max_{1\leqslant i\leqslant n} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} |\boldsymbol{x}_i^T\boldsymbol{\beta}| > c_0\sqrt{\log(p\vee n)}\big) &\leqslant \sum_{i=1}^n P\Big(|\boldsymbol{x}_i^T\boldsymbol{\beta}_0| > \frac{c_0\sqrt{\log(p\vee n)}}{2}\Big) \\ &\quad + \sum_{i=1}^n P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)| > \frac{c_0\sqrt{\log(p\vee n)}}{2}\Big). \end{aligned}$$

Since $\boldsymbol{x}_i^T\boldsymbol{\beta}_0$ is sub-Gaussian with variance proxy $||\boldsymbol{\beta}_0||_2^2\sigma_x^2$, then $P\Big(|\boldsymbol{x}_i^T\boldsymbol{\beta}_0| > \frac{c_0\sqrt{\log(p\vee n)}}{2}\Big) \leqslant 2\exp\Big(-\frac{c_0^2\log(p\vee n)}{8||\boldsymbol{\beta}_0||_2^2\sigma_x^2}\Big)$. In addition,

$$\begin{aligned} P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1} |\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)| > \frac{c_0\sqrt{\log(p\vee n)}}{2}\Big) &\leqslant P\Big(||\boldsymbol{x}_i||_\infty \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} ||(\boldsymbol{\beta}-\boldsymbol{\beta}_0)||_1 > \frac{c_0\sqrt{\log(p\vee n)}}{2}\Big) \\ &\leqslant P\Big(||\boldsymbol{x}_i||_\infty > \frac{c_0\sqrt{\log n\log(p\vee n)}}{2s\big(h^2\vee\sqrt{\frac{\log p}{n}}\big)}\Big) \\ &\leqslant p\exp\Big[-\frac{c\log(p\vee n)}{s^2\big(h^4\vee n^{-1}\log p\big)}\Big], \end{aligned}$$

for positive constant c. Since $s\big(h^2\vee\sqrt{\frac{\log p}{n}}\big) \leqslant d_0$, consider $||\boldsymbol{\beta}_0||_1$ as a constant, we have

$$\begin{aligned} P\big(\max_{1\leqslant i\leqslant n} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1} |\boldsymbol{x}_i^T\boldsymbol{\beta}| > \sqrt{\log(p\vee n)}\big) &\leqslant np\exp\Big(-\frac{c\log(p\vee n)}{s^2\big(h^4\vee n^{-1}\log p\big)}\Big) + 2n\exp\Big(-\frac{\log(p\vee n)}{8||\boldsymbol{\beta}_0||_2^2\sigma_x^2}\Big) \\ &\leqslant \exp[-c_1\log(p\vee n)], \end{aligned}$$

for some universal positive constant c_1. □

Proof of Lemma B.7 The proofs are similar. We only show for (B.1).

Consider the events $\mathcal{H}_n$ and $\mathcal{K}_n$ defined in Lemma B.3. We have

$$\begin{aligned} & P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\big| > t \Big) \\ \leqslant\ & P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\big| > t \,\big|\mathcal{H}_n \cap \mathcal{K}_n\Big) + \exp[-c\log(p\vee n)] \\ \leqslant\ & \sum_{j=1}^{p} P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)\big| > t \,\big|\mathcal{H}_n \cap \mathcal{K}_n\Big) + \exp[-c\log(p\vee n)]. \end{aligned}$$

Write $\boldsymbol{\gamma}(Z_i, \boldsymbol{\beta}, m)$ as $\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)$. Observing that $[h\log(p\vee n)]^{-1}\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)$ satisfies the conditions of Massart's concentration inequality, (Theorem 14.2, Bühlmann and van de Geer (2011)) on $\mathcal{H}_n \cap \mathcal{K}_n$, we have $\forall t > 0$,

$$\begin{aligned} P\Big(& \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big|[nh\log(p\vee n)]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big| \\ & \geqslant \mathrm{E} \quad \Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big|[nh\log(p\vee n)]^{-1} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big|\Big) + t \,\big|\mathcal{H}_n \cap \mathcal{K}_n\Big) \leqslant \exp(-\frac{nt^2}{8}). \end{aligned}$$

Equivalently, $\forall t > 0$,

$$\begin{aligned} P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big| \ \geqslant\ & \mathrm{E}\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big|\Big) \\ & + \sqrt{n}ht\log(p\vee n) \,\big|\mathcal{H}_n \cap \mathcal{K}_n\Big) \leqslant \exp(-\frac{nt^2}{8}) \end{aligned} \tag{B.13}$$

It remains to find an upper bound for $\mathrm{E}\big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} |n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)|\big)$. Let $M_1, \cdots, M_{m(s)}$ denote different subsets of $\{1, \cdots, p\}$, corresponding to different submodels with size at most ks. Note that $m(s) \leqslant \binom{p}{ks}$. Let $S_{M_l} = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 \leqslant c_0 s\max\{h^2, \sqrt{\frac{\log p}{n}}\}, \mathrm{supp}(\boldsymbol{\beta}) = M_l\}$, where $\mathrm{supp}(\boldsymbol{\beta})$ denotes the support set of $\boldsymbol{\beta}$. Then $\mathbb{B}_1 = \cup_{l=1}^{m(s)} S_{M_l}$.

$\forall \delta_n > 0$, (w.l.o.g., $\delta_n \leqslant cs\max\{h^2, \sqrt{\frac{\log p}{n}}\}$), for each S_{M_l}, $(l = 1, \cdots, m(s))$, we can cover it by l_1-balls of radius $c(n)\delta_n$, where $c(n) = \frac{1}{4}[\log(p\vee n)]^{-3/2}$, with centers

$\boldsymbol{\beta}_{l0}^{\circ}, \cdots, \boldsymbol{\beta}_{lN_l}^{\circ}$. Note that this cover can be constructed such that

$$N_l \leqslant \Big(\frac{2c_0 s \max\{h^2, \sqrt{\frac{\log p}{n}}\} + c(n)\delta_n}{c(n)\delta_n}\Big)^{ks}.$$

Denote these N_l balls by $\mathbb{C}(\boldsymbol{\beta}_{ll'}^{\circ})$, $l = 1, \cdots, m(s)$, $l' = 1, \cdots, N_l$.

Observe that $\forall \boldsymbol{\beta} \in \mathbb{B}_1$, $m(\cdot|\boldsymbol{\beta}) \in C_1^1(T)$, where $T = \{t \in \mathbb{R} : |t| \leqslant \sqrt{\log(p \vee n)}\}$. By Theorem 2.7.1 in van der Vaart and Wellner (1996), we have

$$\log N\Big(\frac{\delta_n}{4\log(p \vee n)}, C_1^1(T), ||\cdot||_{\infty}\Big) \leqslant C\frac{\log(p \vee n)}{\delta_n},$$

for positive constant C. So we can find $N_2 \leqslant \exp\big[C\delta_n^{-1}\log(p \vee n)\big]$ brackets of the form $[v_{1a}, v_{2a}]$, $a = 1, \cdots, N_2$, to cover $C_1^1(T)$ such that $\sup_{t\in T}|v_{2a}(t) - v_{1a}(t)| \leqslant \frac{\delta_n}{4\log(p\vee n)}$. Hence $\forall \boldsymbol{\beta} \in \mathbb{B}_1$, $m(\cdot|\boldsymbol{\beta}) \in C_1^1(T)$, we can find l, l' and a such that $\boldsymbol{\beta} \in \mathbb{C}(\boldsymbol{\beta}_{ll'}^{\circ})$ and $v_{1a}(\cdot) \leqslant m(\cdot|\boldsymbol{\beta}) \leqslant v_{2a}(\cdot)$. Let $m_a^{\circ}(\cdot) = \frac{1}{2}[v_{1a}(\cdot) + v_{2a}(\cdot)]$. Conditional on $\mathcal{H}_n \cap \mathcal{K}_n$,

$$\begin{aligned}
&\big|\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) - \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}_{ll'}^{\circ}, m_a^{\circ})\big| \\
\leqslant\ & |m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m_a^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ})| * |\widetilde{\epsilon}_i\boldsymbol{x}_i^T\boldsymbol{\theta}_j| \\
\leqslant\ & \log(p \vee n)\big[\big|m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m_a^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ}|\boldsymbol{\beta})\big| + \big|m(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ}|\boldsymbol{\beta}) - m_a^{\circ}(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^{\circ})\big|\big] \\
\leqslant\ & \log(p \vee n)\big[\sqrt{\log(p \vee n)}c(n)\delta_n + \frac{\delta_n}{4\log(p \vee n)}\big] \leqslant \frac{\delta_n}{2}.
\end{aligned}$$

Then $[\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}_{ll'}^{\circ}, m_a^{\circ}) - \frac{\delta_n}{2}, \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}_{ll'}^{\circ}, m_a^{\circ}) + \frac{\delta_n}{2}]$ forms a δ_n-bracket for $\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)$. Hence, the δ_n-bracket number of the class of functions $\Gamma_j = \{\boldsymbol{\theta}_j^T\boldsymbol{\gamma}(Z, \boldsymbol{\beta}, m) : \boldsymbol{\beta} \in \mathbb{B}_1, m \in \mathbb{M}\}$ is bounded by

$$N_{[]}(\delta_n, \Gamma_j, ||\cdot||_{\infty}) \leqslant c\binom{p}{ks}\Big(\frac{2c_0 s \max\{h^2, \sqrt{\frac{\log p}{n}}\} + c(n)\delta_n}{c(n)\delta_n}\Big)^{ks} \exp\big(C\delta_n^{-1}\log(p \vee n)\big).$$

Then we have

$$\begin{aligned}\log N_{[]}(\delta_n, \Gamma_j, ||\cdot||_\infty) &\leq (ks)\log p + (ks)\log\Big(1 + 8c_0 s h^2 \delta_n^{-1}[\log(p\vee n)]^{3/2}\Big) + \delta_n^{-1}\log(p\vee n) \\ &\leq (ks)\log p + \delta_n^{-1}\log(p\vee n),\end{aligned}$$

since $\log p = o(nh^5)$, $\log(1+b) \leqslant b$ for any $b > 0$, and $sh^2\sqrt{\log(p\vee n)} = o(1)$. It implies that $\log N_{[]}(\delta_n, \Gamma, L_2(P)) \preceq s\log p + \delta_n^{-1}\log(p\vee n)$. Note that $\mathrm{E}|\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)|^2 \leqslant ch^2$, and $|\boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)| \leqslant h\log(p\vee n)$. We have

$$\mathrm{E}\Big[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\big|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big|\Big] \preceq J_{[]}(h, \Gamma_j, L_2(P))\Big(1 + \frac{J_{[]}(h, \Gamma_j, L_2(P))}{h^2\sqrt{n}}h\log(p\vee n)\Big),$$

where

$$\begin{aligned}J_{[]}(h, \Gamma_j, L_2(P)) &= \int_0^h \sqrt{\log N_{[]}(\delta, \Gamma_j, L_2(P))}d\delta \preceq \int_0^h \sqrt{s\log p}d\delta + \int_0^h \sqrt{\log(p\vee n)}\delta^{-1/2}d\delta \\ &\leq h\sqrt{s\log p} + \sqrt{h\log(p\vee n)}.\end{aligned}$$

Note that $J_{[]}(h, \Gamma_j, L_2(P)) = o(1)$, and $\frac{\log(p\vee n)}{\sqrt{n}h} = O(\sqrt{n}h^4) = o(1)$. Consider k as a constant, then the order of h implies that $h\sqrt{s\log p} \leqslant \sqrt{h\log(p\vee n)} * \sqrt{sh} = o\big(\sqrt{h\log(p\vee n)}\big)$. Hence

$$\mathrm{E}\Big[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\big|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big|\Big] \preceq J_{[]}(h, \Gamma_j, L_2(P))[1 + o(1)] \preceq \sqrt{h\log(p\vee n)}.$$

It follows from (B.13) that

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}}\big|n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T\boldsymbol{\gamma}_i(\boldsymbol{\beta}, m)\big| \geqslant c\sqrt{h\log(p\vee n)} + \sqrt{n}ht\log(p\vee n)\,\big|\mathcal{F}_n\cap\mathcal{K}_n\Big) \leqslant \exp(-\frac{nt^2}{8}).$$

Take $t = \sqrt{n^{-1}\log(p \vee n)}$. Recall that $\gamma_1 = \sqrt{h\log(p \vee n)} + h[\log(p \vee n)]^{3/2}$, then

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) \big| \geqslant c\gamma_1 \, \big| \mathcal{H}_n \cap \mathcal{K}_n \Big) \leqslant \exp\big(- \frac{\log(p \vee n)}{8} \big).$$

We have that for certain $c > 0$,

$$P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} \big| n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\gamma}_i(\boldsymbol{\beta}, m) \big| > \gamma_1 \Big) \leqslant \exp - [c\log(p \vee n)].$$

□

Proof of Lemma B.8 The proofs are similar. We only show for (B.3) and (B.5).

Consider the events $\mathcal{G}_n$ and $\mathcal{K}_n$ defined in Lemma B.3. To prove (B.3), we have

$$\begin{aligned}
& P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| > t \Big) \\
\leqslant \; & P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| > t \, \big| \mathcal{G}_n \cap \mathcal{K}_n \Big) + P(\mathcal{G}_n^C) + P(\mathcal{K}_n^C) \\
\leqslant \; & \sum_{j=1}^{p} P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| > t \, \big| \mathcal{G}_n \cap \mathcal{K}_n \Big) + 2\exp - [c\log(p \vee n)].
\end{aligned}$$

Observing that $h^{-2}[\log(p\vee n)]^{-1/2}\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)$ satisfies the conditions of Massart's concentration inequality (Theorem 14.2, Bühlmann and van de Geer (2011)) on $\mathcal{G}_n \cap \mathcal{K}_n$, $\forall t > 0$,

$$\begin{aligned}
P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2} \quad \textstyle\sum_{i=1}^{n} \quad \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| \geqslant \mathrm{E}\big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| \big) \\
+ \quad th^2\sqrt{n\log(p \vee n)} \, \big| \mathcal{G}_n \cap \mathcal{K}_n \Big) \leqslant \exp(-\frac{nt^2}{8}). \qquad \text{(B.14)}
\end{aligned}$$

Then similarly as the proof of Lemma B.7, we can find $N_3 \leqslant \exp\big(C\delta_n^{-1}\sqrt{\log(p \vee n)}\big)$

brackets of the form $[w_{1a}, w_{2a}]$, $a = 1, \cdots, N_3$, to cover $C_1^1(T)$ such that $\sup_{t\in T} |w_{2a}(t) - w_{1a}(t)| \leqslant \frac{\delta_n}{8\sqrt{\log(p\vee n)}}$. Hence $\forall \boldsymbol{\beta} \in \mathbb{B}_1$, $u(\cdot|\boldsymbol{\beta}) \in C_1^1(T)$, we can find $l,\ l'$ and a such that $\boldsymbol{\beta} \in \mathbb{C}(\boldsymbol{\beta}^\circ_{ll'})$ and $w_{1a}(\cdot) \leqslant u(\cdot|\boldsymbol{\beta}) \leqslant w_{2a}(\cdot)$. Let $u_a^\circ(\cdot) = \frac{1}{2}[w_{1a}(\cdot) + w_{2a}(\cdot)]$. We have

$$\begin{aligned}
& |\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u) - \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}^\circ_{ll'}, u_a^\circ)| \\
\leqslant\ & \sqrt{\log(p\vee n)}\big[\big|u(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big| + \big|u(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - u(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ_{ll'}|\boldsymbol{\beta})\big| \\
& + \big|u(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ_{ll'}|\boldsymbol{\beta}) - u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}^\circ_{ll'})\big|\big] \\
\leqslant\ & \sqrt{\log(p\vee n)}\big[\sqrt{\log(p\vee n)}c(n)\delta_n + \frac{2\delta_n}{8\sqrt{\log(p\vee n)}}\big] \leqslant \frac{\delta_n}{2},
\end{aligned}$$

where $c(n) = [4\log(p\vee n)]^{-1}$. So $[\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}^\circ_{ll'}, u_a^\circ) - \frac{\delta_n}{2}, \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}^\circ_{ll'}, u_a^\circ) + \frac{\delta_n}{2}]$ forms a δ_n-bracket for $\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)$. Then we can bound the δ_n-bracket number of the class of functions $V_{1j} = \{\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z, \boldsymbol{\beta}, u) : \boldsymbol{\beta} \in \mathbb{B}_1, u \in \mathbb{U}\}$ by $\log N_{[]}(\delta_n, V_{1j}, L_2(P)) \preceq s\log p + \delta_n^{-1}\sqrt{\log(p\vee n)}$. Note that conditional on $\mathcal{F}_n \cap \mathcal{K}_n$, we have $|\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| \leqslant csh^2\log(p\vee n)$, and $\mathrm{E}\{[\boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)]^2\} \leqslant ch^4$. Hence,

$$\mathrm{E}\Big[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)|\Big] \preceq J_{[]}(h^2, V_{1j}, L_2(P))\Big(1 + \frac{J_{[]}(h^2, V_{1j}, L_2(P))}{h^4\sqrt{n}} sh^2\log(p\vee n)\Big),$$

where $J_{[]}(h^2, V_{1j}, L_2(P)) \preceq h^2\sqrt{s\log p} + [\log(p\vee n)]^{1/4}h \preceq h^{3/4}$, since $ns^2h^6 \to 0$ from the assumption of Theorem 3.2. Hence

$$\mathrm{E}\Big[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)|\Big] \preceq J_{[]}(h^2, V_{1j}, L_2(P))[1 + o(1)] \preceq h^{3/4}.$$

Take $t = \sqrt{n^{-1}\log(p\vee n)}$, then observe that $h^2\log(p\vee n) = O(nh^7) = o(h)$. Hence it follows from (B.14) that

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i, \boldsymbol{\beta}, u)| \geqslant ch^{3/4}\,\big|\mathcal{G}_n \cap \mathcal{K}_n\Big) \leqslant \exp\big(-\frac{\log(p\vee n)}{8}\big).$$

We have that for certain $c > 0$,

$$P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, u\in\mathbb{U}} |n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\theta}_j^T \boldsymbol{\nu}_1(Z_i,\boldsymbol{\beta},u)| > h^{3/4}\Big) \leqslant \exp[-c\log(p\vee n)],$$

which implies (B.3).

Similarly, to prove (B.5), note that according to the assumptions, $||m||_\infty = O(1)$, and conditional on $\mathcal{K}_n$, $|\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})| \leqslant c_0 sh^2\sqrt{\log(p\vee n)}$. Hence, we can find $N_4 \leqslant \exp\big[C\delta_n^{-1}sh^2\log(p\vee n)\big]$ brackets of the form $[v_{1a}, v_{2a}]$, $a = 1,\cdots,N_4$, to cover $C_1^1(T)$ such that $\sup_{t\in T}|v_{2a}(t)-v_{1a}(t)| \leqslant \frac{\delta_n}{4sh^2\log(p\vee n)}$. Hence $\forall\boldsymbol{\beta}\in\mathbb{B}_1$, $m(\cdot|\boldsymbol{\beta})\in C_1^1(T)$, we can find l, l' and a such that $\boldsymbol{\beta}\in\mathbb{C}(\boldsymbol{\beta}_{ll'}^\circ)$ and $v_{1a}(\cdot)\leqslant m(\cdot|\boldsymbol{\beta})\leqslant v_{2a}(\cdot)$. Let $m_a^\circ(\cdot)=\frac{1}{2}[v_{1a}(\cdot)+v_{2a}(\cdot)]$. We have

$$\begin{aligned}
& |\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m) - \boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta}_{ll'}^\circ,m_a^\circ)| \\
\leqslant\ & \sqrt{\log(p\vee n)}\big[\big|m(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta}) - m(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta})\big| * |\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})| \\
& + \big|m(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ|\boldsymbol{\beta}) - m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)\big| * |\boldsymbol{x}_i^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})| \\
& + |m_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_{ll'}^\circ)| * |\boldsymbol{x}_i^T(\boldsymbol{\beta}_{ll'}^\circ-\boldsymbol{\beta})|\big] \\
\leqslant\ & \sqrt{\log(p\vee n)}\big[\frac{sh^2}{4}\delta_n + \sqrt{\log(p\vee n)}c(n)\delta_n + \frac{\delta_n}{4\sqrt{\log(p\vee n)}}\big] \\
=\ & \delta_n\Big[\frac{1}{2}+o(1)\Big],
\end{aligned}$$

since $sh^2 = o(1)$, $c(n) = [4\log(p\vee n)]^{-1}$. Then we can roughly bound the δ_n-bracket number of the class of functions $V_{2j} = \{\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z,\boldsymbol{\beta},m) : \boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}\}$ by $\log N_{[]}(\delta_n, V_{2j}, L_2(P)) \preceq s\log p + \delta_n^{-1}sh^2\sqrt{\log(p\vee n)}$. Similarly, conditional on $\mathcal{G}_n\cap\mathcal{K}_n$, we have that $|\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)| \leqslant csh^2\log(p\vee n)$, and $\mathrm{E}\{[\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)]^2\} \leqslant ch^4$. Hence,

$$\begin{aligned}
\mathrm{E}\Big[\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} |n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\theta}_j^T\boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)|\Big] &\preceq J_{[]}(h^2,V_{2j},L_2(P))\Big(1+\frac{J_{[]}(h^2,V_{2j},L_2(P))}{h^4\sqrt{n}}sh^2\log(p\vee n)\Big) \\
&\preceq h^2\sqrt{s\log p} + [s^2\log(p\vee n)]^{1/4}h^2 \preceq h^{3/2}.
\end{aligned}$$

Take $t = \sqrt{n^{-1}\log(p \vee n)}$, then the Massart's concentration inequality implies that

$$P\Big(\sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} |n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T \boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)| \geqslant ch^{3/2} + h^2\log(p\vee n) \,\big|\, \mathcal{G}_n \cap \mathcal{K}_n \Big) \leqslant \exp(-\frac{\log(p\vee n)}{8}).$$

Note that $h^2\log(p\vee n) = O(h)$. Hence for some $c > 0$,

$$P\Big(\max_{1\leqslant j\leqslant p} \sup_{\boldsymbol{\beta}\in\mathbb{B}_1, m\in\mathbb{M}} |n^{-1/2}\sum_{i=1}^n \boldsymbol{\theta}_j^T \boldsymbol{\nu}_2(Z_i,\boldsymbol{\beta},m)| > h \Big) \leqslant \exp[-c\log(p\vee n)],$$

which implies (B.5). □

Proof of Lemma B.9 The proofs are similar. We only verify (B.8). Consider the event $\mathcal{K}_n$ defined in Lemma B.3. Note that

$$\begin{aligned} & P\Big(\sup_{u\in\mathbb{U}} \big|\big| \sum_{i=1}^n |n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|\big|_\infty > t \Big) \\ \leqslant\ & P\Big(\sup_{u\in\mathbb{U}} \big|\big| n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|\big|_\infty > t \,\big|\, \mathcal{K}_n \Big) + \exp[-c\log(p\vee n)]. \end{aligned}$$

Observing that $h^{-2}[\log(p\vee n)]^{-1/2}\boldsymbol{\psi}(\boldsymbol{x}_i,u)$ satisfies the conditions of Massart's concentration inequality (Theorem 14.2, Bühlmann and van de Geer (2011)) on $\mathcal{K}_n$, we have $\forall t > 0$,

$$\begin{aligned} & P\Big(\sup_{u\in\mathbb{U}} \big|\big| n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|\big|_\infty \geqslant \mathrm{E}\big(\sup_{u\in\mathbb{U}} \big|\big| n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i,u)\big|\big|_\infty \big) + th^2\sqrt{n\log(p\vee n)} \,\big|\, \mathcal{K}_n \Big) \\ \leqslant\ & \exp(-\frac{nt^2}{8}). \end{aligned} \tag{B.15}$$

Then similarly as the proof of Lemma B.7, we can find $N_4 \leqslant \exp\big(C\delta_n^{-1}\sqrt{\log(p\vee n)}\big)$ brackets of the form $[w_{1a}, w_{2a}]$, $a = 1,\cdots,N_3$, to cover $C_1^1(T)$ such that $\sup_{t\in T}|w_{2a}(t) - w_{1a}(t)| \leqslant \frac{\delta_n}{2b\sqrt{\log(p\vee n)}}$. Hence $\forall u(\cdot|\boldsymbol{\beta}_0) \in C_1^1(T)$, we can find a such that $w_{1a}(\cdot) \leqslant$

$u(\cdot|\boldsymbol{\beta}_0) \leqslant w_{2a}(\cdot)$. Let $u_a^\circ(\cdot) = \frac{1}{2}[w_{1a}(\cdot) + w_{2a}(\cdot)]$. Then we have

$$||\boldsymbol{\psi}(\boldsymbol{x}_i, u) - \boldsymbol{\psi}(\boldsymbol{x}_i, u_a^\circ)||_\infty \leqslant b\sqrt{\log(p \vee n)}\big[\big|u(\boldsymbol{x}_i^T\boldsymbol{\beta}_0) - u_a^\circ(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\big| \leqslant \frac{\delta_n}{2}.$$

So $[\boldsymbol{\psi}(\boldsymbol{x}_i, u_a^\circ) - \frac{\delta_n}{2}, \boldsymbol{\psi}(\boldsymbol{x}_i, u_a^\circ) + \frac{\delta_n}{2}]$ forms a δ_n-bracket for $\boldsymbol{\psi}(\boldsymbol{x}_i, u)$. Then we can bound the δ_n-bracket number of the class of functions $\Psi = \{\boldsymbol{\psi}(\boldsymbol{x}, u) : u \in \mathbb{U}\}$ by $\log N_{[]}(\delta_n, \Psi, L_2(P)) \preceq \delta_n^{-1}\sqrt{\log(p \vee n)}$. Conditional on $\mathcal{K}_n$, we haveE$||\boldsymbol{\psi}(\boldsymbol{x}_i, u)\boldsymbol{\psi}(\boldsymbol{x}_i, u)^T||_\infty \leqslant ch^4$, and $||\boldsymbol{\psi}(\boldsymbol{x}_i, u)||_\infty \leqslant ch^2\sqrt{\log(p \vee n)}$. Hence,

$$\mathrm{E}\Big[\sup_{u\in\mathbb{U}}\big|\big|n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i, u)\big|\big|_\infty\Big] \preceq J_{[]}(h^2, \Psi, L_2(P))\Big(1 + \frac{J_{[]}(h^2, \Psi, L_2(P))}{h^4\sqrt{n}}h^2\sqrt{\log(p \vee n)}\Big),$$

where $J_{[]}(h^2, V_1, L_2(P)) \preceq h[\log(p \vee n)]^{1/4} \leqslant c_0(nh^9)^{1/4} = o(h^{3/4})$ for all sufficiently large n. Hence

$$\mathrm{E}\Big[\sup_{u\in\mathbb{U}}\big|\big|n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i, u)\big|\big|_\infty\Big] \preceq h^{3/4}.$$

Take $t = \sqrt{n^{-1}\log(p \vee n)}$. Note that $h^2\log(p \vee n) = o(h)$. It follows from (B.15) that

$$P\Big(\sup_{u\in\mathbb{U}}||n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{x}_i, u)||_\infty \geqslant ch^{3/4}\,\big|\mathcal{K}_n\Big) \leqslant \exp\big[-\frac{\log(p \vee n)}{8}\big],$$

which implies (B.7). □

Proof of Lemma B.10 The proofs are similar. We only show for (B.9). Note that

$$\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1}\big|\widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big| \leqslant \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1}\big|\widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta})\big| + \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1}\big|G(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big|.$$

Assumption (B5) implies that there exists positive constants c_1 and c_2 such that

$$\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1}\big|G(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta}_0)\big| \leqslant \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1} L_1|\boldsymbol{x}^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| \leqslant c_1 sh^2\sqrt{\log(p \vee n)} \leqslant c_2 h.$$

In addition, consider $||\boldsymbol{\beta}_0||_1$ as a constant, then $\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1} |\boldsymbol{x}^T\boldsymbol{\beta}| \leqslant \sqrt{\log(p\vee n)} * (||\boldsymbol{\beta}_0||_1 + c_0 sh^2) \leqslant c_1\sqrt{\log(p\vee n)}$ for positive constant c_1. Hence we can apply Lemma B.4 and derive that

$$P\Big(\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\beta}\in\mathbb{B}_1} \big|\widehat{G}(\boldsymbol{x}^T\boldsymbol{\beta}) - G(\boldsymbol{x}^T\boldsymbol{\beta})\big| \geqslant c_0 sh^2\sqrt{\log(p\vee n)}\Big) \leqslant \exp(-c_1 nh^5).$$

Hence we can conclude (B.9). □

Proof of Lemma B.11 Assumption (B1) indicates that $\inf_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{v} = \xi_1$. Combining with (B5), we can derive that

$$\inf_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v} = \inf_{||\boldsymbol{v}||_2=1} \mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}^T\boldsymbol{\beta}_0)(\boldsymbol{x}^T\boldsymbol{v})]^2\big\} \geqslant a^2 \inf_{||\boldsymbol{v}||_2=1} \boldsymbol{v}^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{v} = a^2\xi_1.$$

Since $\boldsymbol{\Theta} = \boldsymbol{\Omega}^{-1}$, we know that $\boldsymbol{\theta}_j^T\boldsymbol{\Omega}\boldsymbol{\theta}_j = \Theta_{j,j}$. Note that $\tau_{0j}^{-2} = \Theta_{j,j} \leqslant ||\boldsymbol{\theta}_j||_2$, hence $||\boldsymbol{\theta}_j||_2 \geqslant \boldsymbol{\theta}_j^T\boldsymbol{\Omega}\boldsymbol{\theta}_j \geqslant a^2\xi_1||\boldsymbol{\theta}_j||_2^2$, which implies that $\tau_{0j}^{-2} \leqslant ||\boldsymbol{\theta}_j||_2 \leqslant (a^2\xi_1)^{-1} \leqslant M_2/a^2$. Similarly, we can derive that $\tau_{0j}^2 \leqslant \Omega_{j,j} \leqslant ||\boldsymbol{\Omega}||_\infty \leqslant b^2 M_1$. □

Proof of Lemma B.12 It suffices to show that

$$\max_{1\leqslant j,k\leqslant p} \left|\frac{1}{n}\sum_{i=1}^n [\widetilde{Y}_i - \widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right| = o_p(1).$$

Rewrite that

$$
\begin{aligned}
& \max_{1\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}[\widetilde{Y}_i-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right| \\
\leqslant\ & \max_{1\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2\big\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big\}\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k\right| \\
& +\max_{1\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_j^T\boldsymbol{\Lambda}\boldsymbol{\theta}_k\right| \\
& +\max_{1\leqslant j,k\leqslant p}\left|\frac{1}{n}\sum_{i=1}^{n}[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k\right| \\
& +\max_{1\leqslant j,k\leqslant p}\left|\frac{2}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i[G(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-\widehat{G}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})][\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k\right| \\
=\ & \sum_{l=1}^{4}\max_{1\leqslant j,k\leqslant p}|J_{njkl}|,
\end{aligned}
$$

where J_{njkl}'s are defined in the context. To bound $|J_{njk1}|$, consider the event $\mathcal{F}_n = \{\max_{1\leqslant i\leqslant n}|\widetilde{\epsilon}_i| \leqslant (\sigma_\epsilon + M_g)\sqrt{\log(p\vee n)}\}$, which holds with probability at least $1-\exp[-c\log(p\vee n)]$ by the sub-Gaussian property for $\widetilde{\epsilon}$. Lemma B.10 and Lemma B.15 together imply that

$$
\begin{aligned}
\max_{1\leqslant j,k\leqslant p}|J_{njk1}| &\leqslant \max_{1\leqslant i\leqslant n}\big|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big|\Big[\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^6\Big]^{1/3}\max_{1\leqslant j,k\leqslant p}\Big[\frac{1}{n}\sum_{i=1}^{n}|\widehat{\boldsymbol{\theta}}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k|^{3/2}\Big]^{2/3} \\
&\leqslant O_p(h)*\Big[\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^6\Big]^{1/3}*\max_{1\leqslant j\leqslant p}\Big[\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_j|^3\Big]^{2/3} \\
&\leqslant O_p(h)*O_p(1)=O_p(h)=o_p(1),
\end{aligned}
$$

with probability at least $1-\exp[-c_1\log(p\wedge n)]$. Similarly we have $\max_{1\leqslant j,k\leqslant p}|J_{njk3}| \leqslant O_p(s^2h^4)=o_p(1)$, and $\max_{1\leqslant j,k\leqslant p}|J_{njk4}| \leqslant O_p(sh^2)=o_p(1)$. To bound $\max_{1\leqslant j,k\leqslant p}|J_{njk2}|$,

let $\widehat{\mathbf{\Lambda}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\epsilon}_i^2[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i\boldsymbol{x}_i^T$, then we can rewrite it as

$$\begin{aligned}\max_{1\leqslant j,k\leqslant p}|J_{njk2}| &\leqslant \max_{1\leqslant j,k\leqslant p}\left|\boldsymbol{\theta}_j^T\{\widehat{\mathbf{\Lambda}}-\mathbf{\Lambda}\}\boldsymbol{\theta}_k\right| + \max_{1\leqslant j,k\leqslant p}\left|(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)^T\widehat{\mathbf{\Lambda}}\boldsymbol{\theta}_k\right| + \max_{1\leqslant j,k\leqslant p}\left|\widehat{\boldsymbol{\theta}}_j^T\widehat{\mathbf{\Lambda}}(\widehat{\boldsymbol{\theta}}_k-\boldsymbol{\theta}_k)\right| \\ &= \max_{1\leqslant j,k\leqslant p}|J_{njk21}| + \max_{1\leqslant j,k\leqslant p}|J_{njk22}| + \max_{1\leqslant j,k\leqslant p}|J_{njk23}|,\end{aligned}$$

where J_{njk2l}'s are defined clearly. Given $(\epsilon_i, \boldsymbol{x}_i)$, for any $\boldsymbol{\theta}_j$, $\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy at most $b^2[\epsilon_i+g(\boldsymbol{x}_i)]^2(\boldsymbol{x}_i^T\boldsymbol{\theta}_j)^2$. Note that $\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\boldsymbol{x}_i^T\boldsymbol{\theta}_j$ is sub-Gaussian with variance proxy at most $\frac{b^2}{n}\sum_{i=1}^{n}[\epsilon_i+g(\boldsymbol{x}_i)]^2(\boldsymbol{x}_i^T\boldsymbol{\theta}_j)^2 \xrightarrow{p} c$ for constant $c>0$. Lemma B.13 implies that

$$P\Big(\max_{1\leqslant j,k\leqslant p}|J_{njk21}| \geqslant c\sqrt{\frac{\log p}{n}}\Big|\mathcal{F}_n\Big) \leqslant \sum_{j,k} P\Big(|J_{njk21}| \geqslant c\sqrt{\frac{\log p}{n}}\Big|\mathcal{F}_n\Big) \leqslant p^2\exp(-c_1\log p),$$

where $\sqrt{\frac{\log p}{n}} = o(h^{5/2})$ for all sufficiently large n. For J_{njk22}, we can conclude that

$$\begin{aligned}\max_{1\leqslant j,k\leqslant p}|J_{njk22}| &\leqslant b^2\Big[\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^6\Big]^{1/3} * \max_{1\leqslant k\leqslant p}\Big[\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\widehat{\boldsymbol{\theta}}_k|^3\Big]^{1/3} * \max_{1\leqslant j\leqslant p}\Big[\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T(\widehat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j)|^3\Big]^{1/3} \\ &\leqslant O_p(\widetilde{s}^{1/2}h) = o_p(1).\end{aligned}$$

Similar proof works to bound $\max_{1\leqslant j,k\leqslant p}|J_{njk23}|$. The lemma is proved. □

B.6 Auxiliary Results and Lemmas

Definition B.1

A random vector $\boldsymbol{x}\in\mathbb{R}^p$ is said to be sub-Gaussian with variance proxy σ^2 if $\mathrm{E}\boldsymbol{x}=\mathbf{0}$, and for each (fixed) unit vector $\boldsymbol{v}\in\mathbb{R}^p$, its m.g.f satisfies

$$\mathrm{E}[\exp(s\boldsymbol{x}^T\boldsymbol{v})] \leqslant \exp\Big(\frac{\sigma^2 s^2}{2}\Big), \qquad \forall s\in\mathbb{R}.$$

An equivalent definition is that for each (fixed) unit vector $\boldsymbol{v} \in \mathbb{R}^p$ and any $t > 0$, $P(|\boldsymbol{x}^T\boldsymbol{v}| \geqslant t) \leqslant 2\exp\big(-\frac{t^2}{2\sigma^2}\big)$.

Property: $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ are independently sub-Gaussian with variance proxy σ^2, then $\forall t > 0$, $P\big(\max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty > t\big) \leqslant 2np\exp\big(-\frac{t^2}{2\sigma^2}\big)$. As a result,

$$P\Big(\max_{1\leqslant i\leqslant n} ||\boldsymbol{x}_i||_\infty > 2\sigma\sqrt{\log(np)}\Big) \leqslant 2\exp\big[-\log(np)\big].$$

Lemma B.13 (Lemma 14 in Loh and Wainwright (2012))

If $\{\boldsymbol{x}_i \in \mathbb{R}^{p_1} : i = 1, \cdots, n\}$ are independent zero-mean sub-Gaussian random vectors with variance proxy σ_x^2, then for any fixed unit vector $\boldsymbol{v} \in \mathbb{R}^{p_1}$, we have

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n\big[(\boldsymbol{x}_i^T\boldsymbol{v})^2 - \mathrm{E}(\boldsymbol{x}_i^T\boldsymbol{v})^2\big]\Big| \geqslant t\Big) \leqslant 2\exp\big[-cn\min\big(\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\big)\big], \tag{B.16}$$

with a universal constant $c > 0$. Moreover, if $\{\boldsymbol{y}_i \in \mathbb{R}^{p_2} : i = 1, \cdots, n\}$ are independent zero-mean sub-Gaussian random vectors with variance proxy σ_y^2, then

$$P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\big[\boldsymbol{x}_i\boldsymbol{y}_i^T - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{y}_i^T)\big]\Big|\Big|_\infty \geqslant t\Big) \leqslant 6p_1p_2\exp\big[-cn\min\big(\frac{t^2}{\sigma_x^2\sigma_y^2}, \frac{t}{\sigma_x\sigma_y}\big)\big]. \tag{B.17}$$

In particular, if $p = p_1 \vee p_2$ and $\log p = O(n)$, then there are universal positive constants c_0, c_1 and c_2 such that

$$P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^n\big[\boldsymbol{x}_i\boldsymbol{y}_i^T - \mathrm{E}(\boldsymbol{x}_i\boldsymbol{y}_i^T)\big]\Big|\Big|_\infty \geqslant c_0\sigma_x\sigma_y\sqrt{\frac{\log p}{n}}\Big) \leqslant c_1\exp(-c_2\log p). \tag{B.18}$$

□

Lemma B.14 (Lemma 15 in Loh and Wainwright (2012))

Let $\mathbb{K}(s_0) = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}||_2 \leqslant 1, ||\boldsymbol{v}||_0 \leqslant s_0\}$. If $\{\boldsymbol{x}_i \in \mathbb{R}^p : i = 1, \cdots, n\}$ are independent zero-mean sub-Gaussian random vectors with variance proxy σ^2, then there is

a universal constant $c > 0$ such that for any $s_0 \geqslant 1$,

$$P\Big(\sup_{\boldsymbol{v} \in \mathbb{K}(2s_0)} \Big| \frac{1}{n} \sum_{i=1}^{n} \big[(\boldsymbol{x}_i^T \boldsymbol{v})^2 - \mathrm{E}(\boldsymbol{x}_i^T \boldsymbol{v})^2 \big] \Big| \geqslant t \Big) \leqslant 2 \exp\big[- cn \min\big(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2} \big) + 2 s_0 \log p \big]. \quad \square$$

Lemma B.15

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be independent sub-Gaussian random vectors with variance proxy σ^2. For any $s_0 \geqslant 1$, there exists a universal constant $c > 0$ such that for all n sufficiently large.

$$P\Big(\sup_{\boldsymbol{\gamma} \in \mathbb{K}(2s_0)} \Big| \frac{1}{n} \sum_{i=1}^{n} \big[|\boldsymbol{x}_i^T \boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T \boldsymbol{\gamma}|^3 \big] \Big| \geqslant t \Big) \leqslant \exp\big\{ - c \min\big[\frac{nt^2}{\sigma^6}, \frac{(nt)^{2/3}}{\sigma^2} \big] + 2 s_0 \log p \big\}. \quad \square$$

Proof of Lemma B.15 For any fixed $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $||\boldsymbol{\gamma}||_2 \leqslant 1$, $\boldsymbol{x}_i^T \boldsymbol{\gamma}$ is also sub-Gaussian with variance proxy bounded by σ^2. Applying the result on concentration inequality for the polynomial functions of independent sub-Gaussian random variables, that is Theorem 1.4 of Adamczak and Wolff (2015) and the example in their section 3.12, we have $\forall t > 0$, there exist universal positive constants c_1 and c_2 such that

$$P\Big(\frac{1}{n} \sum_{i=1}^{n} \big(|\boldsymbol{x}_i^T \boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T \boldsymbol{\gamma}|^3 \big) \geqslant t \Big) \leqslant c_1 \exp\big\{ - c_2 \min\big[\frac{nt^2}{\sigma^6}, \frac{(nt)^{2/3}}{\sigma^2} \big] \big\}. \qquad \text{(B.19)}$$

Next, we apply the covering technique of Lemma B.14 to extend the above probability bound to uniformly on $\mathbb{K}(s_0) = \{\boldsymbol{v} \in \mathbb{R}^p : ||\boldsymbol{v}||_2 \leqslant 1, ||\boldsymbol{v}||_0 \leqslant s_0\}$. For each subset $\mathcal{U} \subseteq \{1, \cdots, p\}$, define $\mathcal{S}_{\mathcal{U}} = \{\boldsymbol{\gamma} \in \mathbb{R}^p : ||\boldsymbol{\gamma}||_2 \leqslant 1, \mathrm{supp}(\boldsymbol{\gamma}) \subseteq \mathcal{U}\}$. Then $\mathbb{K}(2s_0) = \cup_{|\mathcal{U}| \leqslant 2s_0} \mathcal{S}_{\mathcal{U}}$.

Let $\mathcal{A} = \{u_1, \cdots, u_m\}$ be a $\frac{1}{4}-$cover of $\mathcal{S}_{\mathcal{U}}$, then we can construct $\mathcal{A}$ such that $|\mathcal{A}| \leqslant$

16^{2s_0}. $\forall \boldsymbol{\gamma} \in \mathcal{S}_{\mathcal{U}}$, there is a $\boldsymbol{\xi} \in \mathcal{A}$ such that $||\boldsymbol{\gamma} - \boldsymbol{\xi}||_2 \leqslant 1/4$. Observing that

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - |\boldsymbol{x}_i^T\boldsymbol{\xi}|^3\big)\Big| \;\leqslant\; & \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| + \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}| * |\boldsymbol{x}_i^T\boldsymbol{\xi}| * |\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \\ & + \frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\xi}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})|.\end{aligned}$$

By Hölder inequality,

$$\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \leqslant \Big(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\Big)^{2/3} * \Big(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})|^3\Big)^{1/3}.$$

Since $4(\boldsymbol{\gamma} - \boldsymbol{\xi}) \in \mathcal{S}_{\mathcal{U}}$, we have

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\sup_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^2|\boldsymbol{x}_i^T(\boldsymbol{\gamma}-\boldsymbol{\xi})| \leqslant \frac{1}{4n}\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3.$$

Similarly analysis applies the other two terms. Note that $\max_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 \leqslant \sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3$, then we have

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 \leqslant \max_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 + \frac{3}{4n}\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3,$$

which implies that $\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 \leqslant 4\max_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i^T\boldsymbol{\xi}|^3$. Combining with (B.19) and the union bound, there exists a universal constant $c > 0$ such that

$$\begin{aligned}P\Big(\sup_{\boldsymbol{\gamma}\in\mathcal{S}_{\mathcal{U}}}\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big) \geqslant 4t\Big) \;&\leqslant\; P\Big(\max_{\boldsymbol{\xi}\in\mathcal{A}}\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\xi}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big) \geqslant t\Big) \\ &\leqslant\; 16^{2s_0}\exp\big\{-c\min\big[\frac{nt^2}{\sigma^6}, \frac{(nt)^{2/3}}{\sigma^2}\big]\big\}.\end{aligned}$$

Taking a union bound over the $\binom{p}{2s_0} \leqslant p^{2s_0}$ choices of $\mathcal{U}$ for $\mathbb{K}(2s_0)$ yields that for all n

sufficiently large,

$$P\Big(\sup_{\boldsymbol{\gamma}\in\mathbb{K}(2s_0)} \Big|\frac{1}{n}\sum_{i=1}^{n}\big(|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3 - \mathrm{E}|\boldsymbol{x}_i^T\boldsymbol{\gamma}|^3\big)\Big| \geqslant t\Big) \leqslant \exp\big\{ - c\min\big[\frac{nt^2}{\sigma^6}, \frac{(nt)^{2/3}}{\sigma^2}\big] + 2s_0\log p\big\}.$$

□

Lemma B.16

Under assumptions of Theorem 3.1, there exist universal positive constants c_0, c_1 such that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}} |A_{n1}(t|\boldsymbol{\beta})| \geqslant c_0h^2\Big) \leqslant \exp(-c_1nh^5),$$

where $\mathbb{B}$ and $\mathbb{T}$ are defined as in Lemma 3.1 and Lemma B.4. □

Proof of Lemma B.16 Let $a_{ni}(t|\boldsymbol{\beta}) = K\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)\widetilde{\epsilon}_i$. Then $A_{n1}(t|\boldsymbol{\beta}) = (nh)^{-1}\sum_{i=1}^{n} a_{ni}(t|\boldsymbol{\beta})$. Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. Assumption (B1), (B3) and (B4) together imply that for some constant c large enough,

$$\mathrm{E}[a_{ni}^2(t|\boldsymbol{\beta})] \leqslant (\sigma_\epsilon^2 + M_g^2)\mathrm{E}\Big\{K^2\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)\Big\} = (\sigma_\epsilon^2 + M_g^2)h\int K^2(z)f_{\boldsymbol{\beta}}(t-hz)dz \leqslant ch.$$

Note that $\mathrm{E}a_{ni}(t|\boldsymbol{\beta}) = 0$. Since ϵ_i is sub-Gaussian, $K(\cdot)$ and $g(\cdot)$ are bounded almost everywhere, it is easy to conclude that $\mathrm{E}\big[|a_{ni}(t|\boldsymbol{\beta})|^k\big] \leqslant \frac{1}{2}\mathrm{E}[a_{ni}^2(t|\boldsymbol{\beta})]L^{k-2}k!$, t for some positive real L and every integer $k \geqslant 2$. For any fixed $\boldsymbol{\beta}$ and $0 \leqslant t \leqslant \frac{1}{2L}\sqrt{n\mathrm{E}[a_{ni}^2(t|\boldsymbol{\beta})]}$, by Bernstein's inequality,

$$P\Big(\Big|\sum_{i=1}^{n} a_{ni}(t|\boldsymbol{\beta})\Big| \geqslant 2t\sqrt{cnh}\Big) \leqslant 2\exp(-t^2).$$

Take $t = \sqrt{cnh^5}$. It implies that

$$P\Big(|A_{n1}(t|\boldsymbol{\beta})| \geqslant ch^2\Big) \leqslant 2\exp(-cnh^5).$$

To cover $\mathbb{T}$ with $\delta\sqrt{\log(p \vee n)}-$balls, the covering number $N_1 \leqslant c_1\delta^{-1}$ for sufficiently large c_1. To cover $\mathbb{B}$ with $\delta-$balls, let us consider the unit Euclidean sphere $\mathcal{S}^{ks}$ equipped with the Euclidean metric ρ. It is well known that $N(\delta, \rho, \mathcal{S}^{ks}) \leqslant (1 + \frac{2}{\delta})^{ks}$. Together with the decomposition inspired in **?**,

$$\{\boldsymbol{\beta} \in \mathcal{S}^p : ||\boldsymbol{\beta}||_0 = ks\} = \bigcup_{\mathcal{S}\subseteq[p]:|\mathcal{S}|=ks} \{\boldsymbol{\beta} \in \mathcal{S}^p : \text{supp}(\boldsymbol{\beta}) = \mathcal{S}\},$$

it is easy to show the covering number N_2 satisfies

$$N_2 \leqslant \left\{ \begin{pmatrix} p \\ ks \end{pmatrix} \Big(1 + \frac{2r}{\delta}\Big)^{ks} \right\}^2 \leqslant \left\{ \Big(1 + \frac{2r}{\delta}\Big) \frac{ep}{ks} \right\}^{2ks} \leqslant c_2 \Big(\frac{p}{\delta}\Big)^{2ks},$$

for sufficiently large c_2. Hence to cover $\mathbb{T} \times \mathbb{B}$ with $\delta\sqrt{\log(p \vee n)} \times \delta-$balls, the covering number $N = N_1 * N_2 \leqslant cp^{2ks}\delta^{-(2ks+1)}$. For any $(\boldsymbol{\beta}, t)$ in such a ball with center $(\boldsymbol{\beta}^*, t^*)$, let us take $\delta = \frac{h^3}{2\sqrt{\log(p\vee n)}}$, then conditional on the event $\mathcal{F}_n$, the Lipschitz condition for $K(\cdot)$ implies that

$$\begin{aligned} & \Big|n^{-1}\sum_{i=1}^n [K_h(t - \boldsymbol{x}_i^T\boldsymbol{\beta}) - K_h(t^* - \boldsymbol{x}_i^T\boldsymbol{\beta}^*)]\widetilde{\epsilon}_i\Big| \leqslant c(nh)^{-1}\sum_{i=1}^n \delta\sqrt{\log(p \vee n)}|\widetilde{\epsilon}_i| \\ = \quad & ch^{-1}\delta\sqrt{\log(p \vee n)}n^{-1}\sum_{i=1}^n |\widetilde{\epsilon}_i| \leqslant \frac{ch^2}{2}\sqrt{\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^2}. \end{aligned}$$

Lemma B.2 and Lemma B.13 imply that $P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i^2 - 4(\sigma_\epsilon^2 + M_g^2)\Big| \geqslant \sigma_\epsilon^2 + M_g^2\Big) \leqslant \exp(-cn)$ for some constant $c > 0$. Hence with probability at least $1 - \exp(-cn)$, we have

$\Big|n^{-1}\sum_{i=1}^{n}[K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})-K_h(t^*-\boldsymbol{x}_i^T\boldsymbol{\beta}^*)]\widetilde{\epsilon}_i\Big| \leqslant ch^2/2$. It implies that

$$\begin{aligned}
P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(t|\boldsymbol{\beta})|\geqslant ch^2\Big) &\leqslant P\Big(\bigcup_{(\boldsymbol{\beta},t)\in N}|A_{n1}(t|\boldsymbol{\beta})|\geqslant ch^2/2\Big)\\
&\leqslant \sum_{(\boldsymbol{\beta},t)\in N}P\Big(|A_{n1}(t|\boldsymbol{\beta})|\geqslant ch^2/2\Big)+\exp(-cn)\\
&\leqslant cp^{2ks}\delta^{-(2ks+1)}\exp(-cnh^5)+\exp(-cn)\\
&= c\exp\Big[2ks\log p-(2ks+1)\log\delta-cnh^5\Big]+\exp(-cn).
\end{aligned}$$

Note that $d_0 s\log(p\vee n)\leqslant nh^5$ for constant d_0, we derive that $\delta^{-1}\leqslant 2[\log(p\vee n)]^{-1/10}(n/s)^{3/5}$. Hence we have $-s\log\delta\leqslant s\log(p\vee n)$. It is followed by

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}|A_{n1}(t|\boldsymbol{\beta})|\geqslant c_0h^2\Big)\leqslant\exp(-c_1nh^5),$$

for positive constants c_0 and c_1. We can conclude the lemma. □

Lemma B.17

Under assumptions of Theorem 3.1, there exist universal positive constants c_0, c_1 such that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(t|\boldsymbol{\beta})|\geqslant c_0h^2\Big)\leqslant\exp(-c_1nh^5).$$ □

Proof of Lemma B.17 Let $A_{n2}(t|\boldsymbol{\beta})=(nh)^{-1}\sum_{i=1}^{n}\gamma(z_i)$, where $\gamma(z_i)=K\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)\big(f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(t|\boldsymbol{\beta})\big)$. Observing that $\sup_{t\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{B}}\mathrm{E}[\gamma(z_i)^2]\leqslant c_1h$. Let $f_{\boldsymbol{\beta}}(\cdot)$ denote the p.d.f of $\boldsymbol{x}^T\boldsymbol{\beta}$. Note

that

$$\begin{aligned}
\mathrm{E}\gamma(z_i) &= \mathrm{E}\Big\{K\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)\big(f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(t|\boldsymbol{\beta})\big)\Big\} \\
&= \mathrm{E}\Big\{K\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)\big(G(\boldsymbol{x}_i^T\boldsymbol{\beta}|\boldsymbol{\beta})-G(t|\boldsymbol{\beta})\big)\Big\} \\
&= h\int K(-z)\big(G(t+hz|\boldsymbol{\beta})-G(t|\boldsymbol{\beta})\big)f_{\boldsymbol{\beta}}(t+hz)dz \\
&= h\int K(-z)\Big[G^{(1)}(t|\boldsymbol{\beta})hz+\frac{h^2z^2}{2}G^{(2)}(t_1|\boldsymbol{\beta})\Big]\Big[f_{\boldsymbol{\beta}}(t)+hzf'_{\boldsymbol{\beta}}(\widetilde{t})\Big]dz \\
&= \frac{h^3f_{\boldsymbol{\beta}}(t)}{2}\int z^2K(-z)G^{(2)}(t_1|\boldsymbol{\beta})dz+h^3G^{(1)}(t|\boldsymbol{\beta})\int z^2K(-z)f'_{\boldsymbol{\beta}}(\widetilde{t})dz \\
&\quad +\frac{h^4}{2}\int z^3K(-z)G^{(2)}(t_1|\boldsymbol{\beta})f'_{\boldsymbol{\beta}}(\widetilde{t})dz,
\end{aligned}$$

where t_1 and $\widetilde{t}$ are both between t and $t+hz$. According to (B3)–(B5), we know that $\sup_{t\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{B}} \mathrm{E}\big(A_{n2}(t|\boldsymbol{\beta})\big) \leqslant ch^2$ for some c large enough.

Since $K(\cdot)$ is bounded on the real line, then for any fixed $\boldsymbol{\beta}$ and t, by Bernstein's inequality, $\forall\eta>0$, there exists constant $c>0$ such that

$$P\big(\big|\sum_{i=1}^n\gamma(z_i)-\mathrm{E}\gamma(z_i)\big|\geqslant t\big)\leqslant 2\exp\Big(\frac{-ct^2}{nh}\Big).$$

Note that $\mathrm{E}\gamma(z_i)=O(h^3)$. Take $t=\sqrt{nh^6}$, then we can conclude that $P\Big(\big|A_{n2}(t|\boldsymbol{\beta})\big|\geqslant c_0h^2\Big)\leqslant 2\exp(-c_1nh^5)$ for positive constants c_0 and c_1.

Use $\delta\sqrt{\log(p\vee n)}\times\delta-$balls to cover $\mathbb{T}\times\mathbb{B}$, with the covering number $N\leqslant cp^{2ks}\delta^{-(2ks+1)}$, as shown in the proof of Lemma B.16. For any $(\boldsymbol{\beta},t)$ in such a ball with center $(\boldsymbol{\beta}^*,t^*)$,

we need to bound

$$
\begin{aligned}
& \Big| n^{-1} \sum_{i=1}^{n} \Big[K_h(t - \boldsymbol{x}_i^T \boldsymbol{\beta}) \big(f_0(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G(t|\boldsymbol{\beta}) \big) - K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \big(f_0(\boldsymbol{x}_i^T \boldsymbol{\beta}_0) - G(t^*|\boldsymbol{\beta}^*) \big) \Big] \Big| \\
\leqslant \; & \Big| n^{-1} \sum_{i=1}^{n} \Big[K_h(t - \boldsymbol{x}_i^T \boldsymbol{\beta}) - K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \Big] \big(G(\boldsymbol{x}_i^T \boldsymbol{\beta}_0 | \boldsymbol{\beta}_0) - G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) \big) \Big| \\
& + \Big| n^{-1} \sum_{i=1}^{n} \Big[K_h(t - \boldsymbol{x}_i^T \boldsymbol{\beta}) - K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \Big] \big(G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) - G(t|\boldsymbol{\beta}) \big) \Big| \\
& + \Big| n^{-1} \sum_{i=1}^{n} K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \big(G(t|\boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) + G(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*) - G(t^*|\boldsymbol{\beta}^*) \big) \Big| \\
& + \Big| n^{-1} \sum_{i=1}^{n} K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \big(G(\boldsymbol{x}_i^T \boldsymbol{\beta} | \boldsymbol{\beta}) - G(\boldsymbol{x}_i^T \boldsymbol{\beta}^* | \boldsymbol{\beta}^*) \big) \Big| \\
= \; & \sum_{i=1}^{4} |I_{ni}|,
\end{aligned}
$$

where I_{ni}'s are defined clearly from the context. Take $\delta = \frac{h^3}{4 \log(p \vee n)}$. Then we can derive that

$$
|I_{n1}| \leqslant c(nh)^{-1} \sum_{i=1}^{n} \delta \log(p \vee n) \leqslant c h^{-1} \delta \log(p \vee n) = \frac{ch^2}{4}.
$$

For I_{n2}, the Lipschitz condition for $K(\cdot)$ implies that there exist $\widetilde{t}_i$ between t and $\boldsymbol{x}_i \boldsymbol{\beta}$ such that

$$
\begin{aligned}
|I_{n2}| & \leqslant \Big| n^{-1} \sum_{i=1}^{n} \Big[K_h(t - \boldsymbol{x}_i^T \boldsymbol{\beta}) - K_h(t^* - \boldsymbol{x}_i^T \boldsymbol{\beta}^*) \Big] G^{(1)}(\widetilde{t}_i | \boldsymbol{\beta}) (\boldsymbol{x}_i \boldsymbol{\beta} - t) \Big| \\
& \leqslant c(nh)^{-1} \sum_{i=1}^{n} \delta \log(p \vee n) = \frac{ch^2}{4}.
\end{aligned}
$$

Similarly, we can derive that $|I_{n3}| \leqslant \frac{c_1 h^2}{\sqrt{\log(p \vee n)}}$, and $|I_{n4}| \leqslant \frac{c_1 h^2}{\sqrt{\log(p \vee n)}}$. Combining all the

previous results, we conclude that

$$\Big|n^{-1}\sum_{i=1}^{n}\Big[K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta})\big(f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(t|\boldsymbol{\beta})\big)-K_h(t^*-\boldsymbol{x}_i^T\boldsymbol{\beta}^*)\big(f_0(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)-G(t^*|\boldsymbol{\beta}^*)\big)\Big]\Big| \leqslant ch^2/2,$$

for constant $c > 0$. Then it implies that

$$\begin{aligned} P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}|A_{n2}(t|\boldsymbol{\beta})| \geqslant c_0h^2\Big) &\leqslant P\Big(\bigcup_{(\boldsymbol{\beta},t)\in N}\big|A_{n2}(t|\boldsymbol{\beta})\big| \geqslant c_0h^2/2\Big) \\ &\leqslant \sum_{(\boldsymbol{\beta},t)\in N}P\Big(\big|A_{n2}(t|\boldsymbol{\beta})\big| \geqslant c_0h^2/2\Big) \\ &\leqslant cp^{2ks}\delta^{-(2ks+1)}\exp(-cnh^5) = \exp(-c_1nh^5), \end{aligned}$$

which concludes the lemma. □

Lemma B.18

Under assumptions of Theorem 3.1, there exist universal positive constants c_0, c_1 such that

$$P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}|A_{n3}(t|\boldsymbol{\beta}) - f_{\boldsymbol{\beta}}(t)| \geqslant c_0h^2\Big) \leqslant \exp(-c_1nh^5). \qquad \square$$

Proof of Lemma B.18 Let $A_{n3}(t|\boldsymbol{\beta}) = (nh)^{-1}\sum_{i=1}^{n} z_i(\boldsymbol{\beta};t)$, where $z_i(\boldsymbol{\beta};t) = K\Big(\frac{t-\boldsymbol{x}_i^T\boldsymbol{\beta}}{h}\Big)$. Observing that $\sup_{t\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{B}}\mathrm{E}[z_i^2(\boldsymbol{\beta}|t)] \leqslant c_1h$. We also have

$$\begin{aligned} \mathrm{E}\{z_i(\boldsymbol{\beta};t)\} &= \int K\big(\frac{t-y}{h}\big)f_{\boldsymbol{\beta}}(y)dy = h\int K(-z)f_{\boldsymbol{\beta}}(t+hz)dz \\ &= h\int K(-z)\Big[f_{\boldsymbol{\beta}}(t) + hzf'_{\boldsymbol{\beta}}(t) + \frac{h^2z^2}{2}f''_{\boldsymbol{\beta}}(\widetilde{t})\Big]dz \\ &= hf(t) + \frac{h^3}{2}\int z^2K(-z)f''_{\boldsymbol{\beta}}(\widetilde{t})dz, \end{aligned}$$

where $\widetilde{t}$ is between t and $t+hz$. Assumption (B3)–(B4) imply that $\sup_{t\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{B}} \left|\frac{h^3}{2}\int z^2K(-z)f''_{\boldsymbol{\beta}}(\widetilde{t})dz\right| \leqslant c_0h^3$ for some positive constant c_0.

Since $K(\cdot)$ is bounded on the real line, then for any fixed $\boldsymbol{\beta}$ and t, by Bernstein's inequality, $\forall\eta>0$, there exists constant $c>0$ such that

$$P\Big(\big|\sum_{i=1}^n z_i(\boldsymbol{\beta};t) - \mathrm{E}z_i(\boldsymbol{\beta};t)\big| \geqslant \eta\Big) \leqslant 2\exp\Big(\frac{-c\eta^2}{nh}\Big).$$

Note that $\mathrm{E}\{z_i(\boldsymbol{\beta};t)\} = f(t) + O(h^2)$. Take $\eta = \sqrt{nh^6}$, then we can conclude that $P\Big(\big|A_{n3}(t|\boldsymbol{\beta}) - f(t)\big| \geqslant c_0h^2\Big) \leqslant 2\exp(-c_1nh^5)$ for positive constants c_0 and c_1.

Use $\delta\sqrt{\log p(p\vee n)}\times\delta-$balls to cover $\mathbb{T}\times\mathbb{B}$, with the covering number $N \leqslant cp^{2ks}\delta^{-(2ks+1)}$, as shown in the proof of Lemma B.16. For any $(\boldsymbol{\beta},t)$ in such a ball with center $(\boldsymbol{\beta}^*,t^*)$, let us take $\delta = \frac{h^3}{2\sqrt{\log(p\vee n)}}$, then the Lipschitz condition for $K(\cdot)$ implies that

$$\Big|n^{-1}\sum_{i=1}^n[K_h(t-\boldsymbol{x}_i^T\boldsymbol{\beta}) - K_h(t^*\boldsymbol{x}_i^T\boldsymbol{\beta}^*)]\Big| \leqslant c(nh)^{-1}\sum_{i=1}^n\delta\sqrt{\log(p\vee n)} = ch^2/2.$$

Then it implies that

$$\begin{aligned}
P\Big(\sup_{t\in\mathbb{T},\boldsymbol{\beta}\in\mathbb{B}}\big|A_{n3}(t|\boldsymbol{\beta}) - f(t)\big| \geqslant c_0h^2\Big) &\leqslant P\Big(\bigcup_{(\boldsymbol{\beta},t)\in N}\big|A_{n3}(t|\boldsymbol{\beta}) - f(t)\big| \geqslant c_0h^2/2\Big)\\
&\leqslant \sum_{(\boldsymbol{\beta},t)\in N} P\Big(\big|A_{n3}(t|\boldsymbol{\beta}) - f(t)\big| \geqslant c_0h^2/2\Big)\\
&\leqslant cp^{2ks}\delta^{-(2ks+1)}\exp(-cnh^5) = \exp(-c_1nh^5),
\end{aligned}$$

which concludes the lemma. □

Lemma B.19

Under assumptions (B1) and (B5), if $\log p = O(n)$, then

$$P\Big(\max_{1\leqslant j\leqslant p} \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\Big) \leqslant \exp(-c_2\log p),$$

where $\boldsymbol{\phi}_{0j} = \tau_{0j}^2\boldsymbol{\theta}_j$. □

Proof of Lemma B.19 Note that for any j, $\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}$ and $x_{i,j}$ are both sub-Gaussian with variance proxy no larger than $b^2M_1M_2/a^2\sigma_x^2$ and σ_x^2, by Lemma B.11. Hence the property of sub-Gaussian implies that for any j,k, there exist universal positive constants c_1 and c_2 such that

$$P\Big(\Big|[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}x_{i,k}\Big| \geqslant c_1^2b^2\sigma_x^2t^2\Big) \leqslant P(|\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}| \geqslant c_1\sigma_x t)+P(|\boldsymbol{x}_{i,k}| \geqslant c_1\sigma_x t) \leqslant \exp(-c_2t^2).$$

The definition of $\boldsymbol{d}_{0j}$ implies that $\mathrm{E}\big\{[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\boldsymbol{x}_{i,-j}\big\} = \mathbf{0}$. Hence Theorem 2.2 in Wainwright (2015) implies that for any $j \neq k$, $[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}x_{i,k}$ is sub-Exponential with parameter $c\sigma_x^2$, for constant $c>0$. Hence we obtain

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}x_{i,k}\Big| \geqslant t^2\Big) \leqslant \exp\Big[-cn\min\Big(\frac{t^2}{\sigma_x^4},\frac{t}{\sigma_x^2}\Big)\Big].$$

Take $t = c_1\sigma_x^2\sqrt{\frac{\log p}{n}} = O(\sigma_x^2)$, then we have

$$\begin{aligned}
& P\Big(\max_{1\leqslant j\leqslant p} \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\boldsymbol{x}_{i,-j}\Big|\Big|_\infty \geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\Big) \\
\leqslant\ & \sum_{j=1}^{p}\sum_{k\neq j} P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}x_{i,k}\Big| \geqslant c_1\sigma_x^2\sqrt{\frac{\log p}{n}}\Big) \\
\leqslant\ & p^2\exp(-c\log p) = \exp(-c_2\log p).
\end{aligned}$$

□

Lemma B.20

Assume the conditions of Lemma 3.2 are satisfied, then there exist universal positive constants c_0 and c_1 such that

$$P\Big(\max_{1\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\Big|\Big|_\infty\geqslant c_0h\Big)\leqslant\exp(-c_1\log p),$$

$$P\Big(\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}\boldsymbol{x}_i\boldsymbol{x}_i\Big|\Big|_\infty\geqslant c_0h\Big)\leqslant\exp(-c_1\log p),$$

where $\boldsymbol{\phi}_{0j}=\tau_{0j}^2\boldsymbol{\theta}_j$. □

Proof of Lemma B.20 Proofs are similar. We only show the first one. Note that

$$\begin{aligned}&\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}\Big|\Big|_\infty\\ \leqslant\ &\sup_{1\leqslant i\leqslant n}\big|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big|*\sqrt{\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j})^2}\end{aligned}$$

Let $\mathcal{E}_1=\{\widehat{\boldsymbol{\beta}}\in\mathbb{B}_1,\text{ and }\max_{1\leqslant i\leqslant n}||\boldsymbol{x}_i||_\infty\leqslant\sqrt{\log(p\vee n)}\}$, which holds with probability at leat $1-\exp(-c_1\log p)$, via Theorem 3.1. Then Lemma B.10 and assumption (B5) together implies that

$$\begin{aligned}&P\Big(\sup_{1\leqslant i\leqslant n}\big|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big|\geqslant ch\Big)\\ \leqslant\ &P\Big(\sup_{1\leqslant i\leqslant n}\big|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2-[G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\big|\geqslant ch\,\big|\mathcal{E}_1\Big)+\exp(-c_1\log p)\\ \leqslant\ &n\Big[\exp(-c_1nh^5)+\exp[-c_2\log(p\vee n)]\Big]+\exp(-c_1\log p)\\ =\ &\exp(-d_1\log p),\end{aligned}$$

for universal constant $d_1 > 0$, since $\log(p \vee n) \leqslant d_0 n h^5$. Since $\log p = o(n)$, Lemma B.13 implies that

$$P\Big(\big|\big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\big|\big|_\infty \geqslant 2M_1\Big) \leqslant P\Big(\big|\big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\big|\big|_\infty \geqslant ||\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)||_\infty + d_1\sigma_x^2\sqrt{\frac{\log p}{n}}\Big) \leqslant \exp(-d_2\log p).$$

Since $\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j}$ is sub-Gaussian with variance proxy at most $C\sigma_x^2$ for constant $C > 0$, Lemma B.13 also implies that $P\Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j})^2 \geqslant 2\xi_p M_2^2/a^4\Big) \leqslant \exp(-d_2\log p)$.

Hence we can conclude that

$$\begin{aligned}
& P\Big(\max_{1\leqslant j\leqslant p}\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\{[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2\}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big|\Big|_\infty \geqslant 2chM_2\sqrt{\xi_p M_1}/a^2\Big) \\
\leqslant\ & P\Big(\sup_{1\leqslant i\leqslant n}|[\widehat{G}^{(1)}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})]^2 - [G^{(1)}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)]^2| \geqslant ch\Big) \\
& + P\Big(\big|\big|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\big|\big|_\infty \geqslant 2M_1\Big) + \sum_{j=1}^{p}P\Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{\phi}_{0j})^2 \geqslant 2\xi_p M_2^2/a^4\Big) \\
\leqslant\ & \exp(-c_1\log p).
\end{aligned}$$

□

Appendix C

Proofs of Theorems and Corollaries in Chapter 4

C.1 Proofs of Theorem 4.1

For $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, we have

$$||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}||_\infty = ||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^* - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta}||_\infty \geqslant ||\widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta}||_\infty - ||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^*||_\infty.$$

Hence, on Ω_{n1}, by the definition of $\widehat{\boldsymbol{\beta}}$,

$$\begin{aligned} ||\widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta}||_\infty &\leqslant ||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^*||_\infty + r \\ &\leqslant \varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{\tfrac{\log p}{n}} + r. \end{aligned} \tag{C.1}$$

It follows from Hölder's inequality that

$$\boldsymbol{\delta}^T\widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta} \leqslant ||\widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta}||_\infty||\boldsymbol{\delta}||_1. \tag{C.2}$$

(C.1) and (C.2) together give the upper bound

$$\boldsymbol{\delta}^T\widehat{\boldsymbol{\Gamma}}\boldsymbol{\delta} \leqslant (\varphi(\mathbb{Q}, \sigma_\epsilon)\sqrt{(\log p)/n} + r)||\boldsymbol{\delta}||_1. \tag{C.3}$$

On Ω_{n2}, we have

$$\boldsymbol{\delta}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{\delta} \geqslant \alpha_1||\boldsymbol{\delta}||_2^2 - \tau(n,p)||\boldsymbol{\delta}||_1^2. \tag{C.4}$$

Note that from Lemma 4.1,

$$\begin{aligned} ||\boldsymbol{\delta}||_1 &= ||\boldsymbol{\delta}_T||_1 + ||\boldsymbol{\delta}_{T^C}||_1 \leqslant 2||\boldsymbol{\delta}_T||_1 \\ &\leqslant 2\sqrt{s}||\boldsymbol{\delta}_T||_2 \leqslant 2\sqrt{s}||\boldsymbol{\delta}||_2. \end{aligned}$$

Combining this with (C.3) and (C.4), we have

$$(\alpha_1 - 4\tau(n,p)s)||\boldsymbol{\delta}||_2^2 \leqslant \boldsymbol{\delta}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{\delta} \leqslant 2\sqrt{s}(\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{(\log p)/n} + r)||\boldsymbol{\delta}||_2.$$

This leads to the L_2 estimation error bound

$$\begin{aligned} ||\boldsymbol{\delta}||_2 &\leqslant \frac{2\sqrt{s}}{\alpha_1 - 4\tau(n,p)s}(\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{(\log p)/n} + r) \\ &\leqslant \frac{4\sqrt{s}}{\alpha_1}(\varphi(Q,\sigma_\epsilon)\sqrt{(\log p)/n} + r). \end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} ||\boldsymbol{\delta}||_1 &\leqslant 2\sqrt{s}||\boldsymbol{\delta}||_2 \\ &\leqslant \frac{8s}{\alpha_1}(\varphi(\mathbb{Q},\sigma_\epsilon)\sqrt{(\log p)/n} + r). \end{aligned}$$

C.2 Proofs of Corollary 4.1

Let $\alpha_1 = \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_x)$, $\varphi(\mathbb{Q},\sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, and $\tau(n,p) = d_0\lambda_{\min}(\boldsymbol{\Sigma}_x)\max(\frac{\sigma_z^4}{\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}, 1)\frac{\log p}{n}$, where $\lambda_{\min}(\boldsymbol{\Sigma}_x)$ denotes the smallest eigenvalue of $\boldsymbol{\Sigma}_x$, $c_0 = d_1\sigma_z(\sigma_w + \sigma_\epsilon)$, $\sigma_z^2 = \sigma_x^2 + \sigma_w^2$, and d_0, d_1 are positive constants. For the above choices of α_1, $\tau(n,p)$ and $\varphi(\mathbb{Q},\sigma_\epsilon)$, by

Lemma 1 & Lemma 2 in Loh and Wainwright (2012),

$$P(\Omega_{n1} \cap \Omega_{n2}) \geqslant 1 - c_1 \exp(-c_2 \log p),$$

for some positive constants c_1 and c_2. Corollary 4.1 then follows by applying Theorem 4.1.

C.3 Proofs of Corollary 4.2

Let $\alpha_1 = \frac{1}{2}\lambda_{\min}(\mathbf{\Sigma}_x)$, $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, and $\tau(n,p) = d_0\lambda_{\min}(\mathbf{\Sigma}_x)\max(\frac{\sigma_x^4}{(1-\rho_{\max})^4\lambda_{\min}^2(\mathbf{\Sigma}_x)}, 1)\frac{\log p}{n}$, where $c_0 = d_1\frac{\sigma_x}{1-\rho_{\max}}(\sigma_\epsilon + \frac{\sigma_x}{1-\rho_{\max}})$, and d_0, d_1 are positive constants. For the above choices of α_1, $\tau(n,p)$ and $\varphi(\mathbb{Q}, \sigma_\epsilon)$, by Lemma 3 & Lemma 4 in Loh and Wainwright (2012),

$$P(\Omega_{n1} \cap \Omega_{n2}) \geqslant 1 - c_1 \exp(-c_2 \log p),$$

for some positive constants c_1 and c_2. Corollary 4.2 then follows by applying Theorem 4.1.

C.4 Proofs of Corollary 4.3

Let $\alpha_1 = \frac{1}{2}\lambda_{\min}(\mathbf{\Sigma}_x)$, $\varphi(\mathbb{Q}, \sigma_\epsilon) = c_0||\boldsymbol{\beta}^*||_2$, and $\tau(n,p) = d_0\lambda_{\min}(\mathbf{\Sigma}_x)\max(\frac{K^4\sigma_x^4}{m_{\min}^2\lambda_{\min}^2(\mathbf{\Sigma}_x)}, 1)\frac{\log p}{n}$, where $c_0 = d_1\frac{K\sigma_x}{l_{\min}}(\frac{m_{\min}+Kl_{\min}}{m_{\min}}\sigma_x + \sigma_\epsilon)$, $m_{\min}$ is the minimum element of the covariance matrix $\boldsymbol{M}$, and $l_{\min}$ is the minimum element of the expectation vector $\boldsymbol{l}$, and d_0, d_1 are positive constants.

We will first show that with the above choices of α_1 and $\tau(n,p)$,

$$P(\Omega_{n2}) \geqslant 1 - \eta_1 \exp(-\eta_2 n \min(\frac{m_{\min}^2\lambda_{\min}^2(\mathbf{\Sigma}_x)}{K^4\sigma_x^4}, 1)), \tag{C.5}$$

where η_1 and η_2 are positive constants.

Let $\boldsymbol{\Sigma}_z = \boldsymbol{\Sigma}_x \odot \boldsymbol{M}$. We observe that for any vector $\boldsymbol{\nu} \in \mathbb{R}^p$,

$$\begin{aligned} |\boldsymbol{\nu}^T(\widehat{\boldsymbol{\Gamma}}_{\text{multi}} - \boldsymbol{\Sigma}_x)\boldsymbol{\nu}| &= |\boldsymbol{\nu}^T((\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n} - \boldsymbol{\Sigma}_z) \oslash \boldsymbol{M})\boldsymbol{\nu}| \\ &\leqslant \frac{1}{m_{\min}}|\boldsymbol{\nu}^T(\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n} - \boldsymbol{\Sigma}_z)\boldsymbol{\nu}|. \end{aligned}$$

Since the multiplicative noise $\boldsymbol{U}$ is bounded by K, $\boldsymbol{Z}$ is a sub-Gaussian matrix with parameters $(\boldsymbol{\Sigma}_z, K^2\sigma_x^2)$. So applying Lemma 15 in Loh and Wainwright (2012) with $t = m_{\min}\frac{\lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}$, and define $k = \frac{n}{c\log p}\min(\frac{m_{\min}^2\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}{K^4\sigma_x^4}, 1)$, for some sufficiently small c such that $k \geqslant 1$, similarly as the proof of Lemma 1 in Loh and Wainwright (2012), we can show:

$$\begin{aligned} & P[D(k) \geqslant \frac{\lambda_{\min}(\boldsymbol{\Sigma}_x)}{54}] \\ \leqslant\ & 2\exp\{-d'n\min(\frac{m_{\min}^2\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}{K^4\sigma_x^4}, \frac{m_{\min}\lambda_{\min}(\boldsymbol{\Sigma}_x)}{K^2\sigma_x^2}) + 2k\log p\} \\ \leqslant\ & 2\exp\{-d'n\min(\frac{m_{\min}^2\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}{K^4\sigma_x^4}, 1) + 2k\log p\} \\ \leqslant\ & 2\exp\{-d_2n\min(\frac{m_{\min}^2\lambda_{\min}^2(\boldsymbol{\Sigma}_x)}{K^4\sigma_x^4}, 1)\} \end{aligned}$$

for some positive constant d_2, where

$$D(k) = \sup_{\boldsymbol{\nu}\in\mathbb{K}(2k)} |\boldsymbol{\nu}^T(\widehat{\boldsymbol{\Gamma}}_{\text{multi}} - \boldsymbol{\Sigma}_x)\boldsymbol{\nu}|,$$

and the sparse set:

$$\mathbb{K}(k) = \{\boldsymbol{\nu} :\ ||\boldsymbol{\nu}||_0 \leqslant k, \text{ and } ||\boldsymbol{\nu}||_2 \leqslant 1\}.$$

It then follows from Lemma 13 of Loh and Wainwright (2012) that (C.5) holds.

We next show that with the above choices of $\varphi(\mathbb{Q}, \sigma_\epsilon)$,

$$P(\Omega_{n1}) \geqslant 1 - \eta_3\exp(-\eta_4\log p), \tag{C.6}$$

for some positive constants η_3 and η_4. We have

$$\begin{aligned}
||\widehat{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_x\boldsymbol{\beta}^*||_\infty &= ||\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{y} \oslash \boldsymbol{l} - \mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_i)\boldsymbol{\beta}^*||_\infty \\
&= ||\{\frac{1}{n}\boldsymbol{Z}^T(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) - \mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{x}_i)\boldsymbol{\beta}^*\} \oslash \boldsymbol{l}||_\infty \\
&\leqslant \frac{1}{l_{\min}}||\frac{1}{n}\boldsymbol{Z}^T(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}) - \mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{x}_i)\boldsymbol{\beta}^*||_\infty \\
&\leqslant \frac{1}{l_{\min}}||\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{X}\boldsymbol{\beta}^* - \mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{x}_i^T\boldsymbol{\beta}^*)||_\infty + \frac{1}{l_{\min}}||\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{\epsilon}||_\infty.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
||(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Sigma}_x)\boldsymbol{\beta}^*||_\infty &= ||\{(\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{\Sigma}_z) \oslash \boldsymbol{M}\}\boldsymbol{\beta}^*||_\infty \\
&\leqslant \frac{1}{m_{\min}}||(\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{\Sigma}_z)\boldsymbol{\beta}^*||_\infty.
\end{aligned}$$

Note that $\boldsymbol{X}\boldsymbol{\beta}^*$ is an $n\times 1$ sub-Gaussian matrix with parameters $(\boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}_x\boldsymbol{\beta}^*, ||\boldsymbol{\beta}^*||_2^2\sigma_x^2)$, also $\boldsymbol{Z}\boldsymbol{\beta}^*$ is also an $n\times 1$ sub-Gaussian matrix with parameters $(\boldsymbol{\beta}^{*T}\boldsymbol{\Sigma}_z\boldsymbol{\beta}^*, ||\boldsymbol{\beta}^*||_2^2K^2\sigma_x^2)$. Write $T_1 = \frac{1}{l_{\min}}||\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{X}\boldsymbol{\beta}^* - \mathrm{Cov}(\boldsymbol{z}_i, \boldsymbol{x}_i^T\boldsymbol{\beta}^*)||_\infty$ and $T_2 = \frac{1}{l_{\min}}||\frac{1}{n}\boldsymbol{Z}^T\boldsymbol{\epsilon}||_\infty$. Applying Lemma 14 in Loh and Wainwright (2012), if $n = c'\max\{||\boldsymbol{\beta}^*||_2, 1\}\log p$ with some positive constant c', there exist universal positive constants k_0, k_1 and k_2 such that:

$$\begin{aligned}
P(T_1 > k_0\frac{K\sigma_x^2}{l_{\min}}||\boldsymbol{\beta}^*||_2\sqrt{\frac{\log p}{n}}) &\leqslant k_1\exp(-k_2\log p), \\
P(T_2 > k_0\frac{K\sigma_x\sigma_\epsilon}{l_{\min}}||\boldsymbol{\beta}^*||_2\sqrt{\frac{\log p}{n}}) &\leqslant k_1\exp(-k_2||\boldsymbol{\beta}^*||_2^2\log p),
\end{aligned}$$

and

$$P(||(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Sigma}_x)\boldsymbol{\beta}^*||_\infty > k_0\frac{K^2\sigma_x^2}{m_{\min}}||\boldsymbol{\beta}^*||_2\sqrt{\frac{\log p}{n}}) \leqslant k_1\exp(-k_2\log p).$$

Therefore, there exist universal positive constants d_0, η_3 and η_4 such that the following

inequality is satisfied:

$$\begin{aligned} & P(||\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{\beta}^*||_\infty > d_0 \frac{K\sigma_x}{l_{\min}}(\frac{m_{\min} + Kl_{\min}}{m_{\min}}\sigma_x + \sigma_\epsilon)||\boldsymbol{\beta}^*||_2\sqrt{\frac{\log p}{n}}) \\ \leqslant \quad & \eta_3 \exp(-\eta_4 \log p). \end{aligned}$$

Hence,

$$P(\Omega_{n1} \cap \Omega_{n2}) \geqslant 1 - c_1 \exp(-c_2 \log p), \tag{C.7}$$

for some positive constants c_1 and c_2. Corollary 4.3 then follows by applying Theorem 4.1.