ESSAYS IN ECONOMETRICS AND MACHINE LEARNING

Qingsong Yao

A dissertation (for PhD)

submitted to the Faculty of

the department of Economics

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College

Morrissey College of Arts and Sciences Graduate School

March 2024

ESSAYS IN ECONOMETRICS AND MACHINE LEARNING

Qingsong Yao

Advisors: Shakeeb Khan, Ph.D.

Zhijie Xiao, Ph.D.

Arthur Lewbel, Ph.D.

Abstract

This dissertation consists of three chapters demonstrating how the current econometric problems can be solved by using machine learning techniques. In the first chapter, I propose new approaches to estimating large dimensional monotone index models. This class of models has been popular in the applied and theoretical econometrics literatures as it includes discrete choice, nonparametric transformation, and duration models. A main advantage of my approach is computational. For instance, rank estimation procedures such as those proposed in Han (1987) and Cavanagh and Sherman (1998) that optimize a nonsmooth, non convex objective function are difficult to use with more than a few regressors and so limits their use in with economic data sets. For such monotone index models with increasing dimension, we propose to use a new class of estimators based on batched gradient descent (BGD) involving nonparametric methods such as kernel estimation or sieve estimation, and study their asymptotic properties. The BGD algorithm uses an iterative procedure where the key step exploits a strictly convex objective function, resulting in computational advantages. A contribution of my approach is that the model is large dimensional and semiparametric and so does not require the use of parametric distributional assumptions.

The second chapter studies the estimation of semiparametric monotone index models when the sample size n is extremely large and conventional approaches fail to work due to devastating computational burdens. Motivated by the mini-batch gradient descent algorithm (MBGD) that is widely used as a stochastic optimization tool in the machine learning field, this chapter proposes a novel subsample- and iteration-based estimation procedure. In particular, starting from any initial guess

of the true parameter, the estimator is progressively updated using a sequence of subsamples randomly drawn from the data set whose sample size is much smaller than n. The update is based on the gradient of some well-chosen loss function, where the nonparametric component in the model is replaced with its Nadaraya-Watson kernel estimator that is also constructed based on the random subsamples. The proposed algorithm essentially generalizes MBGD algorithm to the semiparametric setup. Since the new method uses only a subsample to perform Nadaraya-Watson kernel estimation and conduct the update, compared with the full-sample-based iterative method, the new method reduces the computational time by roughly n times if the subsample size and the kernel function are chosen properly, so can be easily applied when the sample size n is large. Moreover, this chapter shows that if averages are further conducted across the estimators produced during iterations, the difference between the average estimator and full-sample-based estimator will be $1/\sqrt{n}$-trivial. Consequently, the averaged estimator is $1/\sqrt{n}$-consistent and asymptotically normally distributed. In other words, the new estimator substantially improves the computational speed, while at the same time maintains the estimation accuracy. Finally, extensive Monte Carlo experiments and real data analysis illustrate the excellent performance of novel algorithm in terms of computational efficiency when the sample size is extremely large.

Finally, the third chapter studies robust inference procedure for treatment effects in panel data with flexible relationship across units via the random forest method. The key contribution of this chapter is twofold. First, it proposes a direct construction of prediction intervals for the treatment effect by exploiting the information of the joint distribution of the cross-sectional units to construct counterfactuals using random forest. In particular, it proposes a Quantile Control Method (QCM) using the Quantile Random Forest (QRF) to accommodate flexible cross-sectional structure as well as high dimensionality. Second, it establishes the asymptotic consistency of QRF under the panel/time series setup with high dimensionality, which is of theoretical interest on its own right. In addition, Monte Carlo simulations are conducted and show that prediction intervals via the QCM have excellent coverage probability for the treatment effects comparing to existing methods in the literature, and are robust to heteroskedasticity, autocorrelation, and various types of model misspecifications. Finally, an empirical application to study the effect of the economic integration between Hong Kong and mainland China on Hong Kong's economy is conducted to highlight the potential of the proposed method.

Contents

1 Estimating High Dimensional Monotone Index Models **1**

1.1 Introduction 1

1.1.1 Notations 3

1.2 The BGD Estimator 4

1.3 Semiparametric BGD Estimation 9

1.3.1 The KBGD Estimator 9

1.3.2 The SBGD Estimator 19

1.4 Monte Carlo Experiments 26

1.5 Empirical Application 29

1.6 Conclusions 30

1.7 Technical Details 31

1.7.1 Lemmas and Proofs 31

1.7.2 Proofs of Theorems 60

2 Stochastic Learning **85**

2.1 Introduction 85

2.1.1 Notations 92

2.2 The Algorithm . . . 92
2.3 Asymptotic Properties of KMBGD Estimator . . . 95
2.4 Monte Carlo Experiments . . . 103
2.4.1 Finite-Sample Performance . . . 105
2.4.2 Computational Efficiency . . . 105
2.5 Real Data Analysis . . . 109
2.5.1 `Run_or_walk_information` . . . 109
2.5.2 `simulated_adult` . . . 111
2.5.3 Revisiting Helpman et al. (2008) . . . 113
2.6 Concluding Remarks . . . 115
2.7 Appendix . . . 116

3 Quantile Control via Random Forest **131**

3.1 Introduction . . . 131
3.1.1 Literature Review . . . 134
3.1.2 Organization . . . 136
3.2 The Model . . . 136
3.3 The Quantile Control Method . . . 138
3.3.1 Prediction Intervals Based on Quantile Regression . . . 139
3.3.2 Quantile Random Forest . . . 141
3.4 Asymptotic Properties of The QCM . . . 145
3.4.1 Concentration of Forest Prediction . . . 146
3.4.2 Bounds on the Bias . . . 149
3.4.3 Consistency . . . 150

3.5 Simulations 155

3.6 Empirical Application 165

3.7 Conclusion 168

3.8 Technical Details 168

3.8.1 Approximating Rectangles 168

3.8.2 Additional Lemmas 170

3.8.3 Proofs of Main Results 171

3.8.4 Proof of Proposition 3.1 181

List of Tables

1.1 Finite Sample Performance of KBGD and SBGD Estimators 27

1.2 Sensitivity of KBGD and SBGD Estimators: Fixed Coefficients 28

1.3 Sensitivity of KBGD and SBGD Estimators: Random Coefficients 28

1.4 Estimation Results 30

2.1 Finite Sample Performance of Kernel-Based Estimators 104

2.2 Comparing Updating Speed 106

2.3 Comparing KMBGD and SBGD Estimators 107

2.4 Comparing True and Estimated Variance 108

3.1 Coverage Probabilities of QCM ($N = 30$) 160

3.2 Comparing Prediction Intervals ($N = 30$, $T = 30$) 162

3.3 Comparing Prediction Intervals ($N = 30$, $T = 100$) 163

List of Figures

2.1 Estimated Coefficients of Walk_or_run_information 110

2.2 ROC of Probit, Logit and KMBGD for Walk_or_run_information 111

2.3 Partial Estimated Coefficients for `Simulated_adult` 112

2.4 Partial Estimated Coefficients for Data in Helpman et al. (2008) 114

3.1 Comparing Prediction Intervals 164

3.2 Actual Outcomes versus Random Forest Prediction 166

3.3 Mean Treatment Effects by Random Forest with 95% CI 167

Chapter 1

Estimating High Dimensional Monotone Index Models by Iterative Convex Optimization

1.1 Introduction

Monotone index models have received a great deal of attention in both the theoretical and applied econometrics literature, as many economic variables of interest are of a limited or qualitative nature. A leading special case in this class is the binary choice model which is usually represented by some variation of the following equation:

$$y_i = I[\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e^{\star} - u_i \geq 0] \tag{1.1}$$

where $I[\cdot]$ is the usual indicator function, y_i is the observed response variable, taking the values 0 or 1 and $\mathbf{X}_{e,i} = \left(X_{0,i}, \mathbf{X}_i^{\mathrm{T}}\right)^{\mathrm{T}}$ is an observed $p+1$ dimensional vector of covariates which effect the behavior of y_i. Both the scalar disturbance term u_i with distribution function denoted by $G(\cdot)$, and the $(p+1)$- dimensional vector $\boldsymbol{\beta}_e^{\star} = \left(\beta^{\star}, \boldsymbol{\beta}^{\star \mathrm{T}}\right)^{\mathrm{T}}$ are unobserved, the latter often being the parameter estimated from a random sample $(y_i, \mathbf{X}_{e,i}), \quad i = 1, 2, ...n$.

The disturbance term u_i is restricted in ways that ensure identification of $\boldsymbol{\beta}_e^{\star}$. Parametric restrictions

specify the distribution of u_i up to a finite dimensional parameter and assume that u_i distributed independently of the covariates $\mathbf{X}_i$. Under such a restriction, $\boldsymbol{\beta}_e^\star$ can be estimated (up to scale) using maximum likelihood or nonlinear least squares. Estimators that are robust to these parametric distributional assumptions have been proposed and analyzed resulting in a variety of estimation procedures for $\boldsymbol{\beta}_e^\star$.

An important class of semiparametric restrictions used in the literature were based on independence/index restrictions. Estimation procedures under this restriction include those proposed by Han (1987), Ichimura (1993), Klein and Spady (1993). These cover but are not limited to the above binary response model. This class of index models have a robustness advantage over parametric approaches, but estimators within this class are difficult to compute[1] due to nonconvexity and in some cases also nonsmoothness of their respective objective functions. For these objective functions, even looking for a local optimum is generally NP-Hard, let alone the global optimum (Murty and Kabadi, 1987). Furthermore the difficulty increases with the dimension of $\mathbf{X}_i$. Recent work which is motivated by computational concerns is Ahn, Ichimura, Powell, and Ruud (2018). However, their two step procedure involves a fully nonparametric estimator in the first stage, so is also not suitable for models with a large number of regressors.

A related drawback of all these procedures is that they are designed to estimate parameters in models of a small and *fixed* dimension. A relatively recent and thriving literature in econometrics and machine learning is recognizing the many advantages of allowing for large dimensional models or models with a large set of controls. This class is a special case of models that consider the situation when the dimension of x_i is large, and this is now often modeled with its dimension increasing with the sample size. Due primarily to its empirical relevance, there has been a burgeoning literature on estimation and inference in certain econometric and statistics models with a large number of regressors or a large number of moment conditions. For a surevey of examples in economics and finance, see Fan et al. (2020). Recent papers include Newey and Windmeijer (2009), Chernozhukov et al. (2017),Belloni et al. (2018), Cattaneo et al. (2018a), Cattaneo et al. (2018b),

Related to our work is the recent literature on estimating large dimensional binary choice or monotone index models in Sur and Candès (2019) and Fan et al. (2020). Sur and Candès (2019) considers inference in a large dimensional logit model, relying on the logistic distribution of the disturbance

[1]Other estimation of index models includes Stoker (1986) and Powell et al. (1989). While these are relatively easy to compute, such derivative based estimators cannot be applied unless all components of $\mathbf{X}_{e,i}$ are continuously distributed.

term where it is shown that χ^2 asymptotic approximations of the LR statistic are suspect when the dimension of x is large. Fan, Han, Li, and Zhou (2020) on the other hand estimate parameters by optimizing the objective function introduced in Han (1987), but with the number parameters increasing with the sample size. Optimizing these rank based objective functions is unfortunately hard even with recent developments in algorithms and search methods for optimizing non smooth and/or non convex objective functions. See for example important recent work based on mixed integer programming (MIP) as in, e.g. Fan et al. (2020) and Shin and Todorov (2021).

Therefore, in light of the drawbacks in the existing literature, this paper proposes a new estimation procedure that is amenable to easier computattion. Specifically we aim to construct a computationally feasible estimator for a semiparametric binary choice and monotone index models with *increasing* dimension based on a convex objective function and then establish its asymptotic properties. As we will discuss in detail in the next section, our algorithm uses an iterative estimator based on a batched gradient descent (BGD) method, and we show how to use nonparametric methods to approximate the distribution in each stage of the iteration. One is the method of sieves[2], and the other is kernel regression.

1.1.1 Notations

Throughout the rest of this paper, to facilitate the description and properties of estimation procedures we will be using the following notation. For any real sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| = 0$, $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$, and $a_n \sim b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$. For any random sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O_p(b_n)$ if for any $0 < \tau < 1$ there are N and $C > 0$ such that $P\{|a_n/b_n| > C\} < \tau$ holds for all $n \geq N$, we write $a_n = o_p(b_n)$ if for any $C > 0$, $\lim_{n\to\infty} P\{|a_n/b_n| > C\} \to 0$. For any Borel sets $A \subseteq \mathbb{R}^k$, denote its Lebesgue measure as $m(A)$. For any symmetric matrix A, we write $A \succ 0$ if A is positive definite, and $A \succeq 0$ if A is positive semi-definite. For any symmetric matrices A and B, we write $A \succ B$ if $A - B \succ 0$ and $A \succeq B$ if $A - B \succeq 0$. For any matrix A, we denote $\sigma(A)$ as its singular value, and denote $\overline{\sigma}(A)$ and $\underline{\sigma}(A)$ as its largest and smallest singular value. For any symmetric matrix A, we denote $\lambda(A)$ as its eigenvalue, and denote $\overline{\lambda}(A)$ and $\underline{\lambda}(A)$ as its largest and smallest eigenvalue. For any vector $\boldsymbol{x} = (x_1, \cdots, x_p)^{\mathrm{T}}$, we denote its Euclidean norm as $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{p} x_i^2}$. For any matrices $A = (a_{ij})_{n\times m}$, we denote $\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$.

[2]See Chen (2007) who pioneered the use of sieve methods in econometrics.

Note that when A is positive semi-definite, there holds $\|A\boldsymbol{x}\| \leq \overline{\lambda}(A) \cdot \|\boldsymbol{x}\|$; for general square matrix A, there holds $\|A\boldsymbol{x}\| \leq \overline{\sigma}(A) \cdot \|\boldsymbol{x}\|$. Finally, for any function $f(\boldsymbol{x})$ with domain D, define $\|f\|_{\infty} = \sup_{\boldsymbol{x} \in D} f(\boldsymbol{x})$.

1.2 The BGD Estimator

To provide some intuition for our semiparametric estimators that will be introduced in the following sections, in this section we consider a simplified version of the model where the cumulative distribution function $G(\cdot)$ is completely known. Under such setup, we explore the *batch gradient descent estimator* (BGD estimator) of $\boldsymbol{\beta}_e^{\star}$ when its dimensionality p may increase, which is also important on its own right. Throughout the following analysis we assume that the data set satisfies the following assumption.

Assumption 1.1. *An i.i.d. data set $\mathscr{D}_n = \{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$ of sample size n is observed, where y_i is generated* [3] *by $y_i = I\left(X_{0,i}\beta_0^{\star} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star} - u_i > 0\right)$ with unobserved shock u_i that is independent of $\mathbf{X}_{e,i}$ and has CDF $G(\cdot)$.*

Given any loss function $\ell_G(\boldsymbol{\beta}_e, \mathbf{X}_e, y)$ that depends on G and is a.s. differentiable with respect to $\boldsymbol{\beta}_e \in \mathcal{B}_e$, the BGD estimator of $\boldsymbol{\beta}_e^{\star}$ is based on the following iteration,

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \partial \ell_G\left(\boldsymbol{\beta}_{e,k}, \mathbf{X}_{e,i}, y_i\right) / \partial \boldsymbol{\beta}_e, \tag{1.2}$$

where $\delta_k > 0$ is the learning rate. Note that $n^{-1}\sum_{i=1}^n \partial \ell_G(\boldsymbol{\beta}_e, \mathbf{X}_{e,i}, y_i)/\partial \boldsymbol{\beta}_e$ constitutes a sample analogue of the derivative $\partial \mathbb{E}\left[\ell_G(\boldsymbol{\beta}_e, \mathbf{X}_e, y)\right]/\partial \boldsymbol{\beta}_e$. Unlike the stochastic gradient descent (SGD) algorithm, in the BGD algorithm, in each round of update we evaluate the derivative of the loss function over all data points. This increases the computational burden but provides a more accurate estimator for the derivative of the expected loss function. Given the initial guess of the parameter, $\boldsymbol{\beta}_{e,1}$, we iterate based on (1.2) until some terminating conditions are reached.

In this paper, we consider the following loss function

$$\ell_G(\boldsymbol{\beta}_e, \mathbf{X}_e, y) = \int_{-A}^{\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e} G(z)\, dz - y\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e, \tag{1.3}$$

[3] Here we are decomposing the vector $\mathbf{X}_{e,i}$ into a scalar component $X_{0,i}$ and the vector $\mathbf{X}_i$, and decomposing the vector of parameters $\boldsymbol{\beta}_e^{\star}$ into the scalar term $\beta_0^{\star}$ and the vector $\boldsymbol{\beta}^{\star}$. As we will see this is done for notational convenience when imposing scale normalizations.

for some sufficiently large positive constant A. The loss function (1.3) was also considered in Agarwal et al. (2014) and has many nice properties. For instance, under some mild conditions, we can show that

$$\begin{aligned}\frac{\partial \mathbb{E}\left(\ell_G\left(\boldsymbol{\beta}_e^\star, \mathbf{X}_e, y\right)\right)}{\partial \boldsymbol{\beta}_e} &= \mathbb{E}\left\{\left(G\left(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e^\star\right) - y\right)\mathbf{X}_e\right\} \\ &= \mathbb{E}\left\{\left(G\left(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e^\star\right) - \mathbb{E}\left(y|\,\mathbf{X}_e\right)\right)\mathbf{X}_e\right\} = 0,\end{aligned}$$

and

$$\frac{\partial^2 \mathbb{E}\left(\ell_G\left(\boldsymbol{\beta}_e, \mathbf{X}_e, y\right)\right)}{\partial \boldsymbol{\beta}_e \partial \boldsymbol{\beta}_e^{\mathrm{T}}} = \mathbb{E}\left\{G'\left(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e\right)\mathbf{X}_e\mathbf{X}_e^{\mathrm{T}}\right\} \succ 0, \forall \boldsymbol{\beta}_e \in \mathcal{B}_e.$$

So $\boldsymbol{\beta}_e^\star$ uniquely minimizes $\mathbb{E}\ell_G\left(\boldsymbol{\beta}_e, \mathbf{X}_e, y\right)$ over $\mathcal{B}_e$. Another desirable property of the loss function (1.3) is that the derivative of (1.3) with respect to $\boldsymbol{\beta}_e$, which is $\left(G\left(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e\right) - y\right)\mathbf{X}_e$, depends only on $G\left(\cdot\right)$ instead of on its derivatives. So when we conduct a semiparametric iteration in the following sections, we only need to nonparametrically approximate $G\left(\cdot\right)$, which is generally more robust compared with approximating its derivatives. Based on loss function (1.3), the BGD estimator is obtained based on the following iteration

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right) - y_i\right)\mathbf{X}_{e,i}. \tag{1.4}$$

We summarize our algorithm as follows in algorithm 1.

Algorithm 1: The BGD Estimator

input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$, sequence of learning rate $\{\delta_k\}_{k=1}^\infty$, initial guess $\boldsymbol{\beta}_{e,1}$, CDF $G\left(\cdot\right)$, and terminating condition $\mathcal{T}$

output: The BGD estimator $\widehat{\boldsymbol{\beta}}_e$

```
k ← 1;
while The terminating condition T is not satisfied do
    β_{e,k+1} ← β_{e,k} − (δ_k/n) Σ_{i=1}^n (G(X_{e,i}^T β_{e,k}) − y_i) X_{e,i};
    k ← k + 1;
β̂_e ← β_{e,k};
```

Remark 1.1. *Key to the above approach is the construction of a convex objective function that facilitates computation even with high dimensions. This transformed convex objective works for any monotone model. In particular, for any model of the form $y_i = G(x_i'\beta) + \epsilon$ with $E[\epsilon_i|x_i] = 0$ and monotone $G(.)$, a similar convex criterion as in (1.3) can be used for inference on β.*

We now describe the asymptotic properties of $\boldsymbol{\beta}_{e,k}$. We first make the following assumption.

Assumption 1.2. *(i) $\mathcal{X}_e = [-1,1]^{p+1}$; (ii) $\mathcal{B}_e$ is convex, and there exists some constant $B_0 > 0$ such that for any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, $|\beta_j| \leq B_0$ for any $0 \leq j \leq p$; (iii) there exists integer υ_G such that G has up to υ_G-th bounded derivatives; (iv) Define $M_n(\boldsymbol{\beta}_e) = \frac{1}{n}\sum_{i=1}^{n} G'\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}$ and $M(\boldsymbol{\beta}_e) = \mathbb{E}[M_n(\boldsymbol{\beta}_e)]$. For any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, there holds $0 < \underline{\lambda}_e \leq \underline{\lambda}(M(\boldsymbol{\beta}_e)) \leq \overline{\lambda}(M(\boldsymbol{\beta}_e)) \leq \overline{\lambda}_e < \infty$.*

Remark 1.2. *Assumption 1.2(i) and Assumption 1.2(ii) are convenient normalizations that facilitate the assessment of our model. Note that to ensure that $\boldsymbol{\beta}_{e,k}$ falls into a compact set for each k, some form of truncation on $\boldsymbol{\beta}_{e,k+1}$ in (1.4) is needed. While according to our results below, as long as $\mathcal{B}_e$ is sufficiently large, it can be shown that $\boldsymbol{\beta}_{e,k}$ will fall into $\mathcal{B}_e$ for all k with probability going to 1. We then assume that $\boldsymbol{\beta}_{e,k} \in \mathcal{B}_e$ for all k. Assumption 1.2(iii) imposes some smoothness conditions on G, where the requirement on υ_G will be stated in the following propositions and theorems. Assumption 1.2(iv) requires that the eigenvalue of $M_n(\boldsymbol{\beta}_e)$ is bounded from both below and above uniformly over $\mathcal{B}_e$.*

For any $\boldsymbol{\beta}_e \in \mathcal{B}_e$, define $\Delta\boldsymbol{\beta}_e = \boldsymbol{\beta}_e - \boldsymbol{\beta}_e^{\star}$. Also define $\varepsilon_i = y_i - G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right)$, where $\mathbb{E}\left[\varepsilon_i \middle| \mathbf{X}_{e,i}\right] = 0$. When Assumption 1.1 and Assumption 1.2 hold, we have the following result.

Theorem 1.1. *Suppose that Assumption 1.1 and Assumption 1.2 hold with $\upsilon_G = 3$, that $p^5(\log p)^2 n^{-1} \to 0$, that the learning rate is chosen such that $\delta_k = \delta \leq 2/\left(3\overline{\lambda}_e\right)$, and that $\boldsymbol{\beta}_e$ is updated based on algorithm 1. We have that*

(i) Define

$$k_{1,n}^{BGD} = \frac{\log\left\|\Delta\boldsymbol{\beta}_{e,1}\right\| + \frac{1}{2}\log\left(n/\left(p\log p\right)\right)}{-\log\left(1-\underline{\lambda}_e\delta/2\right)},$$

we then have

$$\sup_{k \geq k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right);$$

(ii) Define $k_{2,n}^{BGD}$ such that $\left(1-\underline{\lambda}_e\delta\right)^{k_{2,n}^{BGD}}\sqrt{p\log p} \to 0$, we have

$$\sup_{k \geq k_{2,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}} - M^{-1}(\boldsymbol{\beta}_e^{\star})\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| = o_p\left(1/\sqrt{n}\right);$$

(iii) For any $k \geq k_{1,n}^{BGD} + k_{2,n}^{BGD} + 1$, define $\widehat{\boldsymbol{\beta}}_e = \widehat{\boldsymbol{\beta}}_k$. Also define

$$\Sigma_1^{\star} = M^{-1}(\boldsymbol{\beta}_e^{\star})\,\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]M^{-1}(\boldsymbol{\beta}_e^{\star}),$$

and

$$\widehat{\Sigma}_{1,n} = M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right\}M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right),$$

where $G_i^\star = G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^\star\right)$ *and* $\widehat{G}_i = G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_e\right)$. *Suppose further that* $\mathbb{E}\left(\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)$ *has uniformly (with respect to* p*) upper bounded eigenvalues, there holds*

$$\left\|\widehat{\Sigma}_{1,n} - \Sigma_1^\star\right\| \to_p 0.$$

(iv) For any $p+1$ *vector* ρ *such that* $\lim_{n\to\infty}\|\rho\| < \infty$, $\lim_{n\to\infty}\rho^{\mathrm{T}}\Sigma_1^\star\rho = \sigma^2(\rho)$, *and that* $\rho^{\mathrm{T}}M^{-1}\left(\boldsymbol{\beta}_e^\star\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i} \to_d N\left(0,\sigma^2(\rho)\right)$, *we have that*

$$\rho^{\mathrm{T}}\Delta\widehat{\boldsymbol{\beta}}_e/\sqrt{\widehat{\sigma}^2(\rho)/n} \to_d N(0,1),$$

where $\widehat{\sigma}^2(\rho) = \rho^{\mathrm{T}}\widehat{\Sigma}_{1,n}\rho$.

Proof of Theorem 1.1. See subsection 1.7.2. □

When p is fixed, Theorem 1.1(i) implies that $\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| = O_p\left(1/\sqrt{n}\right)$, and Theorem 1.1(ii) implies that for k sufficiently large, the BGD estimator is an asymptotically linear estimator, so there holds $\sqrt{n}\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}} \to_d N\left(0,\Sigma_1^\star\right)$ by the central limit theorem. The asymptotic variance can be estimated based on Theorem 1.1(iii). The number of iterations required to obtain root-n consistency, $k_{1,n}^{BGD}$, is determined by many factors including the sample size n, the distance between the true parameter and the initial guess $||\Delta\boldsymbol{\beta}_{e,1}||$, as well as the lower bound of the eigenvalues of $M_n\left(\boldsymbol{\beta}_e\right)$. In general, $k_{1,n}^{BGD}$ is of order $O(\log n)$, but in practice when we apply the above algorithm, the specific number of iteration is difficult to determine. For detailed discussion of the number of iterations, see Remark 1.5 at the end of Section 1.4. The inference on $\boldsymbol{\beta}_e^\star$ based on the BGD estimator is given by Theorem 1.1(iv). Note that for any given vector ρ, we require that $\frac{1}{\sqrt{n}}\rho^{\mathrm{T}}M^{-1}\left(\boldsymbol{\beta}_e^\star\right)\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}$ is asymptotically normally distributed. An alternative approach is to apply the high-dimensional central limit theorem to $\frac{1}{n}\sum_{i=1}^{n}M^{-1}\left(\boldsymbol{\beta}_e^\star\right)\mathbf{X}_{e,i}\varepsilon_i$ (e.g., Chernozhukov et al., 2017).

Before we conclude this section and move to semiparametric estimation, we further comment on Theorem 1.1. Different from the stochastic gradient descent algorithm (e.g., Toulis and Airoldi, 2017), we show in Theorem 1.1 that the learning rate δ_k can be selected as a sufficiently small constant. Indeed, in the following results, we show that δ_k can decay to zero at any rate as long

as $\sum_{k=1}^{\infty}\delta_k = \infty$ holds, and the choice of δ_k will not change the asymptotic results displayed in Theorem 1.1. In particular, we have the following proposition.

Theorem 1.2. *Suppose that all the conditions in Theorem 1.1 hold and that $\boldsymbol{\beta}_e$ is updated based on algorithm 1. For any sequence of tuning parameters $\{\delta_k\}_{k=1}^{\infty}$ satisfying $\delta_k \geq 0$, $\delta_k \to 0$, $\limsup_{k\to\infty}\delta_{k-1}/\delta_k < \infty$, and $\sum_{k=1}^{\infty}\delta_k = \infty$, we have that*

(i) Define $\widetilde{k}_{1,n}^{BGD}$ such that $\sum_{k=1}^{\widetilde{k}_{1,n}^{BGD}}\delta_k \geq \underline{\lambda}_e^{-1}\left\{\log\left(n/p\left(\log p\right)\right) + 2\log\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|\right\}$, and that $\sup_{k\geq\widetilde{k}_{1,n}^{BGD}+1}\delta_k \leq 2/\underline{\lambda}_e$, then there holds

$$\sup_{k\geq\widetilde{k}_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right);$$

(ii) Define $\widetilde{k}_{2,n}^{BGD}$ such that $\sum_{k=\widetilde{k}_{1,n}^{BGD}+1}^{k=\widetilde{k}_{2,n}^{BGD}}\delta_k/\log p \to \infty$, then we have that

$$\sup_{k\geq\widetilde{k}_{2,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k+\widetilde{k}_{1,n}^{BGD}} - M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| = o_p\left(1/\sqrt{n}\right);$$

(iii) For any $k \geq \widetilde{k}_{1,n}^{BGD} + \widetilde{k}_{2,n}^{BGD} + 1$, define $\widehat{\boldsymbol{\beta}}_e = \widehat{\boldsymbol{\beta}}_k$. We have that Theorem 1.1(iii) and (iv) hold.

Proof of Theorem 1.2. See subsection 1.7.2. □

Theorem 1.2 shows that the choice of the learning rate basically does not affect the convergence rate as well as the asymptotic distribution of the BGD estimators. The main advantage of using a sequence of decaying learning rates is that we do not need to choose the constant δ as required in Theorem 1.1, since for k sufficiently large, $\delta_k \leq 2/\left(3\overline{\lambda}_e\right)$ will automatically hold. However, the disadvantage of using decaying learning rates is that such procedure takes much longer time to converge because the magnitude of the update in the k-th round decreases as k increases. For instance, suppose that we choose $\delta_k \sim k^{-\upsilon}$ for some $0 \leq \upsilon < 1$, we have that $\sum_{j=1}^{k}\delta_j \sim k^{1-\upsilon}$. Then to ensure that $\sum_{j=1}^{\widetilde{k}_{1,n}^{BGD}}\delta_j \geq \underline{\lambda}_e^{-1}\left(\log n + 2\log\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|\right)$, we need $\widetilde{k}_{1,n}^{BGD} \sim (\log n)^{\frac{1}{1-\upsilon}}$. Obviously, setting $\upsilon = 0$ leads to $k \sim \log n$, which corresponds to the requirement in Theorem 1.1(i); when $\upsilon > 0$, we can see that more rounds of iteration is needed compared with required in Theorem 1.1(i).

1.3 Semiparametric BGD Estimation

In the previous section, we focused on iterative estimators based on the BGD algorithm for the parametric binary choice models. We show that when the CDF of the error term is known, the iterative estimators based on the BGD algorithm are consistent and attain asymptotic normality under mild conditions. However, having prior knowledge of the form of G is generally too strong an assumption. In most applications, the source of the individual shock u in Assumption 1.1 is difficult to justify, which makes it quite difficult, if not completely impossible, to know the exact expression of G. In this scenario, the algorithm proposed in the previous section is infeasible. To overcome such problem, this section generalizes the BGD estimator proposed in Section 1.2 to the semiparametric setting where G is unknown.

In this setup, to ensure identification we set $\beta_0^\star$ to be 1, so our estimation target is $\boldsymbol{\beta}^\star$. To simplify our notation, we denote the space of $\mathbf{X}$ as $\mathcal{X}$, and the corresponding parameter space of $\boldsymbol{\beta}$ as $\mathcal{B}$. Suppose that an initial guess for $\boldsymbol{\beta}^\star$ is given by $\boldsymbol{\beta}_1$. In the k-th round of iteration, to update $\boldsymbol{\beta}$ based on the BGD algorithm, we require the knowledge of G as in Section 1.2, which is infeasible when G is unknown. A natural idea is that we can construct an estimator for G based on the index constructed from the updated parameter in the previous round. More intuitively, suppose for a moment that in the k-th round of iteration, $\boldsymbol{\beta}_k$ happens to be identical to the unknown true parameter $\boldsymbol{\beta}^\star$, then we have that $G(z) = \mathbb{E}\left[y|\, X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^\star = z\right] = \mathbb{E}\left[y|\, X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_k = z\right]$ for any $z \in R$.

This motivates semiparametric estimation by using nonparametric methods to estimate $G(\cdot)$. We consider kernel estimation and the method of sieves in each of the following subsections.

1.3.1 The KBGD Estimator

In this section we consider tkernel estimation to estimate $G(\cdot)$. The Nadaraya-Watson kernel estimator of $G(\cdot)$ is of the form

$$\widehat{G}(z|\,\boldsymbol{\beta}_k) = \frac{\sum_{j=1}^n K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_j}{\sum_{j=1}^n K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right)}, z \in R, \tag{1.5}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is some kernel function, and h_n is some bandwidth parameter depending on n. Given the estimated CDF $\widehat{G}(\cdot|\,\boldsymbol{\beta}_k)$, we can update the parameter as if it were the

true CDF $G(\cdot)$. In particular, $\boldsymbol{\beta}_k$ is updated as

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_i. \tag{1.6}$$

Keep updating $\boldsymbol{\beta}_k$ based on (1.5) and (1.6), until some terminating conditions are reached. The resulting estimator is labeled as the *kernel-based batch gradient descent estimator* (KBGD estimator). We summarize our algorithm as follows in algorithm 2.

Algorithm 2: The KBGD Estimator

input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^{n}$, sequence of learning rate $\{\delta_k\}_{k=1}^{\infty}$, initial guess $\boldsymbol{\beta}_1$, kernel function K, bandwidth h_n, and terminating condition $\mathcal{T}$

output: The KBGD estimator $\widehat{\boldsymbol{\beta}}$

$k \leftarrow 1$;
while *The terminating condition* $\mathcal{T}$ *is not satisfied* **do**
  for $i \leftarrow 1$ **to** n **do**
    $\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k \right) \leftarrow \frac{\sum_{j=1}^{n} K_{h_n}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}} \boldsymbol{\beta}_k \right) y_j}{\sum_{j=1}^{n} K_{h_n}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}} \boldsymbol{\beta}_k \right)}$;
  $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_{e,i}$;
  $k \leftarrow k + 1$;
$\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_k$;

Remark 1.3. *In essence, the KBGD estimator can not be classified as a BGD estimator based on a semiparametric loss function. In the semiparametric setup, given any loss function* $\ell_G(\boldsymbol{\beta}, \mathbf{X}_e, y)$ *(quadratic distance in Ichimura (1993), log-likelihood in Klein and Spady (1993), or loss function given in (1.3)) with unknown function* G*, it's a common practice to replace* G *with its nonparametric estimator* $\widehat{G}$ *and then minimize (or maximize) the resulting loss function to obtain the estimator of* $\boldsymbol{\beta}$*. Note that under the single-index framework,* $\widehat{G}$ *usually involves the unknown parameter* $\boldsymbol{\beta}$*, which is of the form* $\widehat{G}(\cdot) = \widehat{G}(\cdot | \boldsymbol{\beta})$*. In this scenario, the BGD estimator is constructed by the following iteration*

$$\boldsymbol{\beta}_{k+1}^{BGD} = \boldsymbol{\beta}_k^{BGD} - \frac{\delta_k}{n} \sum_{i=1}^{n} \frac{\partial \ell_{\widehat{G}\left(\cdot | \boldsymbol{\beta}_k^{BGD}\right)} \left(\boldsymbol{\beta}_k^{BGD}, \mathbf{X}_{e,i}, y_i \right)}{\partial \boldsymbol{\beta}},$$

where $\partial \ell_{\widehat{G}\left(\cdot | \boldsymbol{\beta}_k^{BGD}\right)} \left(\boldsymbol{\beta}_k^{BGD}, \mathbf{X}_{e,i}, y_i \right) / \partial \boldsymbol{\beta}$ *involves* $\partial \widehat{G}(\cdot | \boldsymbol{\beta}_k) / \partial \boldsymbol{\beta}$*, a complicated functions of* $\boldsymbol{\beta}_k$*. In particular, the BGD estimator under loss function (2.7) is given by*

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k \right) + \int_{-\infty}^{X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k} \frac{\partial \widehat{G}(z | \boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}} dz - y_i \right) \mathbf{X}_i.$$

Obviously, an additional term is introduced compared with (1.6). On the contrary, during the construction (1.6), we take G *as given when taking the first order derivative of the loss function and*

then replace the unknown G with its non-parametric estimator in the derivative. More specifically, the KBGD estimator is updated as follows

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} \frac{\partial \ell_G \left(\boldsymbol{\beta}_k, \mathbf{X}_{e,i}, y_i \right)}{\partial \boldsymbol{\beta}} \Big|_{G(\cdot)=\widehat{G}(\cdot|\boldsymbol{\beta}_k)},$$

so additional terms involving $\partial \widehat{G}\left(\cdot|\,\boldsymbol{\beta}_k\right)/\partial \boldsymbol{\beta}$ are avoided. Finally, as we discussed in Section 1.2, the derivative of loss function (1.3) with respect to $\boldsymbol{\beta}$ depends only on G, so we also avoid approximating the derivative of G, which has poorer finite-sample performance compared with approximating G. Such update also ensures contraction map under some conditions, see **??**.

For any fixed z and $\boldsymbol{\beta}$, under mild conditions there holds $\widehat{G}\left(z|\,\boldsymbol{\beta}\right) \to_p \mathbb{E}\left[y|\, X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = z\right]$. Denote such limit as $L\left(z, \boldsymbol{\beta}\right)$. Obviously, $L\left(z, \boldsymbol{\beta}^\star\right) = G\left(z\right)$ holds for any $z \in \mathbb{R}$. Before we move to a formal description of the statistical properties of the KBGD estimator based on (1.6), we first provide some further discussion on $L\left(z, \boldsymbol{\beta}\right)$. For simplicity, in the following we only focus on the case where all the covariates are continuous which permit continuous joint density function. We leave further discussion of the case where some covariates are discrete to Remark 1.6. We point that when there are discrete covariates, our algorithm can be directly applied without any modification, although some further assumptions will be required.

When all the covariates are continuous, denote the joint density of $\mathbf{X}_e$ and $\mathbf{X}$ as $f_e\left(\mathbf{X}_e\right) = f_e\left(X_0, \mathbf{X}\right)$ and $f\left(\mathbf{X}\right) = \int f_e\left(X_0, \mathbf{X}\right) dX_0$, respectively. Denote $z\left(\mathbf{X}_e, \boldsymbol{\beta}\right) = X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. Also denote $f_{\mathbf{X},z}\left(\mathbf{X}, z|\,\boldsymbol{\beta}\right)$ as the joint density of $\mathbf{X}$ and $z\left(\mathbf{X}_e, \boldsymbol{\beta}\right)$ given $\boldsymbol{\beta}$. Note that for any $\boldsymbol{x}$ and z,

$$\begin{aligned} P\left[\mathbf{X} \le \boldsymbol{x}, z\left(\mathbf{X}_e, \boldsymbol{\beta}\right) \le z\right] &= \int_{\widetilde{\mathbf{X}} \le \boldsymbol{x}, \widetilde{X}_0 + \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta} \le z} f_e\left(\widetilde{X}_0, \widetilde{\mathbf{X}}\right) d\widetilde{X}_0 d\widetilde{\mathbf{X}} \\ &= \int_{\widetilde{\mathbf{X}} \le \boldsymbol{x}} \left[\int_{\widetilde{X}_0 \le z - \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}} f_e\left(\widetilde{X}_0, \widetilde{\mathbf{X}}\right) d\widetilde{X}_0 \right] d\widetilde{\mathbf{X}}. \end{aligned}$$

This implies that the joint density of $\mathbf{X}$ and $z\left(\mathbf{X}_e, \boldsymbol{\beta}\right)$ given $\boldsymbol{\beta}$ is given by

$$f_{\mathbf{X},z}\left(\mathbf{X}, z|\,\boldsymbol{\beta}\right) = f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right), \tag{1.7}$$

and the marginal density of $z\left(\mathbf{X}_e, \boldsymbol{\beta}\right)$ is given by

$$f_z\left(z|\,\boldsymbol{\beta}\right) = \int_{\mathcal{X}} f_{\mathbf{X},z}\left(\mathbf{X}, z|\,\boldsymbol{\beta}\right) d\mathbf{X} = \int_{\mathcal{X}} f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right) d\mathbf{X}. \tag{1.8}$$

Define $f_{\mathbf{X}|z}(\mathbf{X}|z,\boldsymbol{\beta}) = f_{\mathbf{X},z}(\mathbf{X},z|\boldsymbol{\beta})/f_z(z|\boldsymbol{\beta})$ as the conditional density of $\mathbf{X}$ given z and $\boldsymbol{\beta}$, we have that

$$\begin{aligned} L(z,\boldsymbol{\beta}) &= \mathbb{E}\left(G\left(z-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\middle| z(\mathbf{X}_e,\boldsymbol{\beta})=z\right) \\ &= \int_{\mathcal{X}} G\left(z-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right) f_{\mathbf{X}|z}(\mathbf{X}|z,\boldsymbol{\beta})\,d\mathbf{X}, \end{aligned} \tag{1.9}$$

where $\Delta\boldsymbol{\beta} = \boldsymbol{\beta}-\boldsymbol{\beta}^{\star}$.

Based on the above notations, now we formally study the asymptotic properties of the KBGD estimator under increasing dimensions. We first introduce some further assumptions.

Assumption 1.3. *The kernel function $K(\cdot)$ satisfies: (i) K is bounded and twice continuously differentiable with bounded first and second derivatives, and the second derivative satisfies Lipschitz condition on the whole real line; (ii) $\int K(s)\,ds = 1$; (iii) there exists positive integer υ_K such that $\int s^{\upsilon}K(s)\,du = 0$ for $1 \le \upsilon \le \upsilon_K - 1$ and $\int u^{\upsilon_K}K(u)\,du \neq 0$; (iv) $K(s) = 0$ for $|s| > 1$.*

Assumption 1.4. *(i) There exists some constant $\zeta > 1$ such that $\zeta^{-1} \le f_e(\mathbf{X}_e) \le \zeta$ holds for all $\mathbf{X}_e \in \mathcal{X}_e$; (ii) there exists positive integer υ_f such that $f_e(\mathbf{X}_e)$ has bounded up to υ_f-th derivatives.*

Remark 1.4. *Assumption 1.4(i) together with Assumption 1.2(i) is a commonly-used assumption in the machine learning literature (e.g., Wager and Athey, 2018). It basically requires that the joint density of $\mathbf{X}_e$ is uniformly bounded from both above and below over $\mathcal{X}_e$, so the density does not degenerate over $\mathcal{X}_e$. Assumption 1.4(i) basically allows us to construct a subset of $\mathcal{X}_e$ such that $f_z(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta})$ is uniformly lowered bounded from zero over such subset.*

The following lemma will be useful in the proof of our theorem.

Lemma 1.1. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), Assumption 1.3, and Assumption 1.4 hold with $\upsilon_G = 3$, $\upsilon_K = 2$, and $\upsilon_f = 3$. Define $\psi(n,p,h) = h^{-1}\sqrt{\log(pnh^{-1})/n} + h^2$. If $h_n \to 0$ and $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}(n,p,h_n) \to 0$ further hold, we have that*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}(z(\mathbf{X}_{e,i},\boldsymbol{\beta})|\boldsymbol{\beta})\mathbf{X}_i - \mathbb{E}\left[L(z(\mathbf{X}_{e,i},\boldsymbol{\beta}),\boldsymbol{\beta})\mathbf{X}_i\right]\right\| = O_p\left(p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}(n,p,h_n)\right).$$

Proof of Lemma 1.1. See subsection 1.7.1. □

Lemma 1.1 implies that $\frac{1}{n}\sum_{i=1}^{n}\widehat{G}(Z(\mathbf{X}_{e,i},\boldsymbol{\beta})|\boldsymbol{\beta})\mathbf{X}_i$ will be closer to $\mathbb{E}\left[L(z(\mathbf{X}_{e,i},\boldsymbol{\beta}),\boldsymbol{\beta})\mathbf{X}_i\right]$ uniformly with respect to $\boldsymbol{\beta}$ as n increases. Note that such uniform convergence results are free of

trimming; we do not need to trim $\mathbf{X}_{e,i}$ even when the density of $z(\mathbf{X}_{e,i},\boldsymbol{\beta})$ is small. So even when $\widehat{G}(z(\mathbf{X}_{e,i},\boldsymbol{\beta})|\boldsymbol{\beta})$ is a poor estimator for $L(z(\mathbf{X}_{e,i},\boldsymbol{\beta}),\boldsymbol{\beta})$ for some $\mathbf{X}_{e,i}$ and $\boldsymbol{\beta}$, our results are still valid. While on the same time, the cost of not conducting any trimming is that our guaranteed convergence rate depends heavily on the dimensionality. As is required in Lemma 1.1, the dimension p must satisfy $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}(n,p,h_n)\to 0$. Suppose that $p/n\to 0$ and we choose $h_n=((\log n)/n)^{1/6}$, we have that $\psi(n,p,h_n)\sim((\log n)/n)^{1/3}$. This implies that when p is fixed, the convergence rate in Lemma 1.1 is $((\log n)/n)^{1/3(p+1)}$. When p increases with n, the dimension p should satisfy $p\log p=O(\log n)$, implying that p is allowed to increase only mildly with n. The restriction on p basically comes from the fact that as $\mathbf{X}_{e,i}$ moves towards the boundary of $\mathcal{X}_e$, the density of random variable $z(\mathbf{X}_{e,i},\boldsymbol{\beta})$ decreases faster towards zero given a larger p, which makes the convergence rate sensitive to the increase of p.

For notational simplicity, in the following we denote $z(\mathbf{X}_{e,i},\boldsymbol{\beta}_k)$ and $z(\mathbf{X}_{e,i},\boldsymbol{\beta}^\star)$ as $z_{i,k}$ and $z_i^\star$. Based on the results in Lemma 1.1, we have that under all conditions as imposed in Lemma 1.1, there holds

$$\boldsymbol{\beta}_{k+1}=\boldsymbol{\beta}_k-\delta_k\mathbb{E}\left[\left(L(z_{i,k},\boldsymbol{\beta}_k)-G(z_i^\star)\right)\cdot\mathbf{X}_i\right]+\delta_k\cdot(\text{small order terms}). \tag{1.10}$$

Note that $z_{i,k}=z_i^\star+\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_k$ and $L(z_{i,k},\boldsymbol{\beta}_k)=\int_{\mathcal{X}}G\left(z_{i,k}-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)f_{\mathbf{X}|z}(\mathbf{X}|z_{i,k},\boldsymbol{\beta}_k)\,d\mathbf{X}$, so $(L(z_{i,k},\boldsymbol{\beta}_k)-G(z_i^\star))\cdot\mathbf{X}_i$ equals to

$$\begin{aligned}&\left\{\int_{\mathcal{X}}\left[G\left(z_i^\star+\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_k-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)-G(z_i^\star)\right]f_{\mathbf{X}|z}(\mathbf{X}|z_{i,k},\boldsymbol{\beta}_k)\,d\mathbf{X}\right\}\cdot\mathbf{X}_i\\&=\int_0^1\int_{\mathcal{X}}\left[G'\left(z_i^\star+t(\mathbf{X}_i-\mathbf{X})^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)f_{\mathbf{X}|z}(\mathbf{X}|z_{i,k},\boldsymbol{\beta}_k)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathbf{X}_i\mathbf{X}^{\mathrm{T}}\right)\right]\Delta\boldsymbol{\beta}_k d\mathbf{X}dt,\end{aligned} \tag{1.11}$$

where the integration is understood to be element-wise. To further simplify our notation, define

$$W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)=G'\left(z(\mathbf{X}_e,\boldsymbol{\beta}^\star)+\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\boldsymbol{X}|z}\left(\widetilde{\mathbf{X}},\Big|\,z(\mathbf{X}_e,\boldsymbol{\beta}),\boldsymbol{\beta}\right),$$

$$V\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)=\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathbf{X}\widetilde{\mathbf{X}}^{\mathrm{T}}\right)W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right),$$

and

$$\Lambda(\boldsymbol{\beta})=\mathbb{E}\left[\int_{\mathcal{X}}V(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta})\,d\mathbf{X}\right],$$

we have that

$$\mathbb{E}\left[\left(L\left(z_{i,k},\boldsymbol{\beta}_k\right)-G\left(z_i^\star\right)\right)\cdot\mathbf{X}_i\right]=\int_0^1 \Lambda\left(\boldsymbol{\beta}^\star+t\Delta\boldsymbol{\beta}_k\right)\Delta\boldsymbol{\beta}_k dt,$$

which indicates that

$$\Delta\boldsymbol{\beta}_{k+1}=\left\{\int_0^1\left(I_p-\delta_k\Lambda\left(\boldsymbol{\beta}^\star+t\Delta\boldsymbol{\beta}_k\right)\right)dt\right\}\Delta\boldsymbol{\beta}_k+\delta_k\cdot(\text{small order terms}).$$

To ensure that with probability going to 1 the above iteration shrinks $\|\Delta\boldsymbol{\beta}_k\|$, we make the following assumption.

Assumption 1.5. *There hold*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Lambda\left(\boldsymbol{\beta}\right)+\Lambda^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\right)\leq\overline{\lambda}_\Lambda<\infty,$$

and

$$\inf_{\boldsymbol{\beta}\in\mathcal{B}}\underline{\lambda}\left(\Lambda\left(\boldsymbol{\beta}\right)+\Lambda^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\right)\geq\underline{\lambda}_\Lambda>0.$$

Based on the above assumptions, we have the following result.

Theorem 1.3. *Suppose that Assumption 1.1, Assumption 1.2(i)–(iii), Assumption 1.3–**??** hold with $\upsilon_G=3$, $\upsilon_K=2$, and $\upsilon_f=3$, $\delta_k=\delta$ such that $\delta<\min\left\{1/\left(2\underline{\lambda}_\Lambda\right),1/\left(4p^2\left\|G'\right\|_\infty\right)\right\}$, and that $\boldsymbol{\beta}$ is updated based on algorithm 2. Define*

$$k_{1,n}^{KBGD}=\frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right)-\log\left(p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\right)}{-\log\left(1-\delta\underline{\lambda}_\Lambda/4\right)}.$$

Then if $h_n\to0$ and $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\to0$ hold, we have that

$$\sup_{k\geq k_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\right).$$

In particular, if h_n is chosen such that $h_n=\left(\left(\log n\right)/n\right)^{1/6}$, then

$$\sup_{k\geq k_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(p^{\frac{5p+1}{2(p+1)}}\left(\frac{\log n}{n}\right)^{\frac{1}{3p+3}}\right).$$

Proof of Theorem 1.3. See subsection 1.7.2. □

Theorem 1.3 implies that the iterative estimator based on (1.5) and (1.6) is consistent under in-

creasing dimensions, no matter whether the starting point is close to the unknown true parameter or not. However, the convergence speed heavily depends on the dimensionality of the problem, p, even when p is fixed. This is not ideal under our single-index setup but is not surprising since our algorithm does not involve any trimming procedure.

We proceed to establish the asymptotic normality of the KBGD estimator. Due to technical difficulties, throughout the following analysis in this section we only consider the case where p is fixed. As we can see in Theorem 1.3, even in the case of fixed dimensionality, the guaranteed convergence rate of the KBGD estimator based on (1.5) and (1.6) is at best $\left(\left(\log n\right)/n\right)^{\frac{1}{3p+3}}$, which still depends on p. To obtain asymptotic normality, we need to slightly modify our algorithm to get rid of the dependence on dimensionality. In particular, we introduce trimming to our algorithm. When updating the parameter, we only use observations that fall into a pre-selected region as did in Ichimura (1993). In particular, the algorithm is modified as,

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} I_i^{\phi} \cdot \left(\widehat{G}\left(z_{i,k} | \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_i, \tag{1.12}$$

where $\widehat{G}\left(z_{i,k} | \boldsymbol{\beta}_k \right) = \widehat{G}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k \right) | \boldsymbol{\beta}_k \right)$ is defined in (1.5), $I_i^{\phi} = I\left(\mathbf{X}_{e,i} \in \mathcal{X}_e^{\phi} \right)$, and $\mathcal{X}_e^{\phi}$ is a subset of $\mathcal{X}_e$ given by

$$\mathcal{X}_e^{\phi} = \{ \mathbf{X}_e \in \mathcal{X}_e : |X_j| \leq 1 - \phi, 0 \leq j \leq p \} \tag{1.13}$$

for some $\phi > 0$ whose value will be determined later. Different from (1.6), the update of $\boldsymbol{\beta}_k$ based on (1.12) uses only a subset of the whole sample for which the covariate vector $\mathbf{X}_{e,i}$ falls into $\mathcal{X}_e^{\phi}$. The reason why we choose the trimming set as in (1.13) is that, as we show in the subsection 3.8.1, for any $0 < \phi < 1$, there holds $\inf_{(\mathbf{X}_e, \boldsymbol{\beta}) \in \mathcal{X}_e^{\phi} \times \mathcal{B}} f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) \geq C \phi^p p^{-p}$ for some constant $C > 0$ that depends on ϕ. When p and ϕ are both fixed, $f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right)$ is uniformly lower bounded from zero for any combination $(\mathbf{X}_e, \boldsymbol{\beta}) \in \mathcal{X}_e^{\phi} \times \mathcal{B}$, so the uniform estimation accuracy of $L\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right)$ over $\mathbf{X}_{e,i}$ and $\boldsymbol{\beta}$ will be improved. Note that trimming will cause some efficiency loss by dropping some observations, but such loss can be controlled to be small if we choose ϕ to be close to zero. We also point that trimming is only applied to the update of the parameter; when nonparametrically estimating G, we still use all the data points.

To simplify our following notation, given the trimming parameter ϕ, we denote $I^{\phi} \cdot \mathbf{X}$ as $\mathbf{X}^{\phi}$. We also define

$$\Lambda_{\phi}\left(\boldsymbol{\beta} \right) = \mathbb{E}\left[I_i^{\phi} \cdot \int_{\mathcal{X}} V\left(\mathbf{X}_{e,i}, \mathbf{X}_e, \boldsymbol{\beta} \right) d\mathbf{X} \right].$$

The following theorem provides a counterpart to the results in Theorem 1.3.

Theorem 1.4. *Suppose that all the assumptions and conditions on υ_G, υ_K, and υ_f in Theorem 1.3 hold. Suppose moreover that $h_n \to 0$, $\delta_k = \delta < \min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right), 1/\left(4p^2 \left\|G'\right\|_{\infty}\right)\right\}$, $\phi < \delta\underline{\lambda}_{\Lambda}/\left(16p^2 \left\|G'\right\|_{\infty} \zeta\right)$, and that $\boldsymbol{\beta}$ is updated under (1.5) and (1.12) (the trimmed version of algorithm 2). Define*

$$\widetilde{k}_{1,n}^{KBGD} = \frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right) - \log\left(\psi\left(n,p,h_n\right)\right)}{-\log\left(1-\delta\underline{\lambda}_{\Lambda}/8\right)},$$

then there holds

$$\sup_{k \geq \widetilde{k}_{1,n}^{KBGD}+1} \left\|\Delta\boldsymbol{\beta}_k\right\| = O_p\left(\psi\left(n,p,h_n\right)\right).$$

Proof of Theorem 1.4. See subsection 1.7.2. □

Note that when p is fixed, $\psi\left(n,p,h_n\right)$ no longer depends on p asymptotically. The improvement over the convergence rate basically comes from the improvement of the uniform convergence rate of the kernel estimator due to trimming. Also note that under trimming, the minimum number of iteration in Theorem 1.3(i), $\widetilde{k}_{1,n}^{KBGD}$, is of order $\log n$ as long as $nh_n \to \infty$. This implies that under trimming, a faster convergence rate is guaranteed with the minimum number of iterations being of the same magnitude as that of the estimator without trimming.

We now proceed to establish the asymptotic normality of $\boldsymbol{\beta}_k$. Define

$$\boldsymbol{\xi}_n^{\phi} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_i^{\star}\middle|\boldsymbol{\beta}^{\star}\right) - y_i\right)\mathbf{X}_i^{\phi}.$$

We note that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+1} &= \Delta\boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^{\phi}, \\
&= \Delta\boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) - \widehat{G}\left(z_i^{\star}\middle|\boldsymbol{\beta}^{\star}\right)\right)\mathbf{X}_i^{\phi} - \delta_k\boldsymbol{\xi}_n^{\phi} \\
&= \int_0^1 \left\{ I_p - \frac{\delta_k}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^{\phi}\left.\frac{\partial\widehat{G}\left(z\left(\boldsymbol{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right]\right\} dt\Delta\boldsymbol{\beta}_k - \delta_k\boldsymbol{\xi}_n^{\phi}, \qquad (1.14)
\end{aligned}$$

where the integration is understood to be element-wise. To understand the properties of the above algorithm, we need the following lemmas.

Lemma 1.2. *Suppose that all the assumptions in Theorem 1.3 hold with $\upsilon_G = 4$, $\upsilon_K = 3$, and*

$\upsilon_f = 4$. *For any sequence of subset* $\{\mathcal{B}_n\}_{n=1}^{\infty}$ *with* $\mathcal{B}_n \subseteq \mathcal{B}$, *we have that*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}-\Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right\|=O_p\left(h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n}+h_n^3+\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\|\Delta\boldsymbol{\beta}\|\right).$$

Proof of Lemma 1.2. See subsection 1.7.1. □

Lemma 1.3. *Suppose that all the assumptions in Theorem 1.3 hold with* $\upsilon_G = 4$, $\upsilon_K = 3$, *and* $\upsilon_f = 4$. *If* h_n *is chosen such that* $h_n^6 n \to 0$, *we have that* $\sqrt{n}\boldsymbol{\xi}_n^{\phi} \to_d N\left(0,\Sigma_{\boldsymbol{\xi}}^{\phi}\right)$, *where*

$$\Sigma_{\boldsymbol{\xi}}^{\phi}=\mathbb{E}\left[\left(1-G\left(z_i^{\star}\right)\right)G\left(z_i^{\star}\right)\left(\mathbf{X}_i^{\phi}-\mathbb{E}\left(\mathbf{X}_i^{\phi}\middle|\,z_i^{\star}\right)\right)\left(\mathbf{X}_i^{\phi}-\mathbb{E}\left(\mathbf{X}_i^{\phi}\middle|\,z_i^{\star}\right)\right)^{\mathrm{T}}\right].$$

Proof of Lemma 1.3. See subsection 1.7.1. □

Now we are in a position to illustrate the results of the asymptotic normality of our KBGD estimator.

Theorem 1.5. *Suppose that all the assumptions in Theorem 1.3 hold with* $\upsilon_G = 4$, $\upsilon_K = 3$, *and* $\upsilon_f = 4$. *Suppose moreover that* $\delta_k = \delta < \min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right), 1/\left(4p^2\left\|G'\right\|_{\infty}\right)\right\}$, $\phi < \delta\underline{\lambda}_{\Lambda}/\left(16p^2\left\|G'\right\|_{\infty}\zeta\right)$, h_n *is chosen such that* $nh_n^6 \to 0$ *and* $h_n^4 n/\left(\log n\right)^2 \to \infty$, *and that* $\boldsymbol{\beta}$ *is updated under (1.5) and (1.12). Then*

(i) There holds

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+k_{2,n}^{KBGD}+1}\|\Delta\boldsymbol{\beta}_k\|=O_p\left(n^{-1/2}\right),$$

where $k_{2,n}^{KBGD}$ *is given by*

$$k_{2,n}^{KBGD}=\frac{\log\left(n^{1/2}\right)+\log\left(\psi\left(n,p,h_n\right)\right)}{-\log\left(1-\delta\underline{\lambda}_{\Lambda}/16\right)};$$

(ii) Define $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}_k$ *for any* $k-\widetilde{k}_{1,n}^{KBGD}-k_{2,n}^{KGBD}\to\infty$, *we have that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\star}\right)\to N\left(0,\Sigma_{\boldsymbol{\beta}}^{\phi}\right),$$

where $\Sigma_{\boldsymbol{\beta}}^{\phi}=\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\Sigma_{\boldsymbol{\xi}}^{\phi}\left(\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}$.

Proof of Lemma 2.3. See subsection 1.7.2. □

We introduce the estimator for the variance matrix, based on which the confidence interval of $\boldsymbol{\beta}^{\star}$

can be then constructed.

Theorem 1.6. *Suppose that all the assumptions and conditions in Theorem 1.5 hold. Suppose also that $\widehat{\boldsymbol{\beta}}$ is defined as in Theorem 1.5. Define*

$$\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi}=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)^{\mathrm{T}}\right),$$

and

$$\widehat{\Lambda}_{\phi}\left(\widehat{\boldsymbol{\beta}}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\widehat{\boldsymbol{\beta}}\right)\middle|\widehat{\boldsymbol{\beta}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}},$$

where

$$\widehat{G}_i=\frac{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)y_j}{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)},\ \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)=\frac{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)\mathbf{X}_j^{\phi}}{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)},$$

and $\widehat{z}_i=X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. Then we have that

$$\left\|\widehat{\Lambda}_{\phi}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi}\left(\widehat{\Lambda}_{\phi}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}-\Sigma_{\boldsymbol{\beta}}^{\phi}\right\|\rightarrow_p 0.$$

Proof of Theorem 1.6. See subsection 1.7.2. □

We finally provide some remarks for the KBGD estimators.

Remark 1.5. *We first provide some remarks on the implementation of our KBGD estimator. The KBGD estimator might be sensitive to the data magnitude. So when implementing such an estimator, we recommend first standardizing the data so that each covariate has zero mean and unit variance. Note that when constructing the KBGD estimator, we normalize the coefficient of $X_{0,i}$ to 1, indicating that the coefficients of $\mathbf{X}_{e,i}$ can not all be zeros. So we need to test whether at least one covariate affects the conditional probability of $y_i=1$. One option is to run a Logit or Probit regression and test whether all the coefficients are equal to zero.*

When applying our algorithm, it is also crucial to determine the learning rate δ, bandwidth of kernel estimator h_n, and terminating conditions of the algorithm. In Theorem 1.5, the tuning parameter δ is required to be smaller than $1/\left(2\underline{\lambda}_{\Lambda}\right)$ and $1/\left(4p^2\left\|G'\right\|_{\infty}\right)$, neither of which is known. So we recommend setting δ to be 1 in the first place, and gradually shrink it if the iteration does not converge. For the choice of the bandwidth h_n, Lemma 2.3 requires that h_n is chosen such that $nh_n^6\rightarrow 0$ and $nh_n^4/\left(\log n\right)^2\rightarrow\infty$. As a rule of thumb, we recommend choosing $h_n=C\cdot n^{-1/5}$. For the choice of the constant C, we can choose $C=C_k=std(z_{i,k})$ for the k-th round of iteration

and $C = std(\widehat{z}_i)$ *when estimating the variance* $\Sigma_{\boldsymbol{\beta}}^{\phi}$. *We finally discuss the terminating conditions. As we show in Theorem 1.5, to obtain root-n consistency and asymptotic normality, the iteration number is required to be only of order* $\log(n)$. *However, such rule can not be directly applied to determine the number of iterations since the initial distance* $\|\Delta\boldsymbol{\beta}_1\|$ *as well as the lower bounded on the eigenvalues* $\underline{\lambda}_{\Lambda}$ *are both unknown. We recommend the terminating condition* $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}| < \varrho$ *for some predetermined tolerance* ϱ. *During the simulation, we choose* $\varrho = 10^{-5}$. *Note that in many cases,* $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}|$ *may not be monotonically decreasing with* k; *in some extreme cases,* $\max_{1\leq j\leq p}|\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}|$ *may even be oscillating and does not shrink to zero. On these condition, we recommend decreasing* δ *or choosing* $h_n = C \cdot n^{-1/5}$ *with* $C = 1$ *when iterating. If the maximum distance still oscillates, we recommend stop iteration when the maximum distance achieves its minimum value.*

Remark 1.6. *Our previous discussion has be confined to the case where all the covariates are continuously distributed, while our algorithm can be directly applied to the case where there are discrete covariates without any modifications. The basic reason is that, in contrast to the average derivative approach (Stoker, 1986; Powell et al., 1989) that uses the differentiation with respect to covariates, the KBGD estimator performs differentiation with respect to the parameters, so it does not impose requirements on the continuity of the covariates. It should be noted that we do require at least one continuous covariate to guarantee identification of the parameters. For simplicity, we recommend choosing a continuous covariate as the standardization covariate* X_0. *Finally, we point out that stronger assumption should be imposed to make our results valid when there are discrete covariates. In particular, suppose that* $\mathbf{X}_e = \left(\mathbf{X}_c^{\mathrm{T}}, \mathbf{X}_d^{\mathrm{T}}\right)^{\mathrm{T}}$, *where* $\mathbf{X}_c$ *is the collection of all the continuous covariates, whereas* $\mathbf{X}_d$ *is the collection of all the discrete covariates. Also denote the density function of* $\mathbf{X}_c$ *conditional on* $\mathbf{X}_d$ *as* $f_{\mathbf{X}_c|\mathbf{X}_d}(\mathbf{X}_c|\mathbf{X}_d)$. *Then we require that all the conditions imposed on the* $f_e(\mathbf{X}_e)$ *hold for* $f_{\mathbf{X}_c|\mathbf{X}_d}(\mathbf{X}_c|\mathbf{X}_d)$ *for any realizations of* $\mathbf{X}_d$.

1.3.2 The SBGD Estimator

In the previous section, we introduced the KBGD algorithm, where the update of the parameter is based on a BGD-type procedure while the unknown CDF is replaced with its Nadaraya-Watson kernel estimator constructed by the initial parameter. In this section, we consider an alternative nonparametric approximation for the unknown CDF based on the method of sieves. Given a set of basis functions $\{r_j(z)\}_{j=0}^{\infty}$ that is complete in $C(\mathbb{R})$ space, any smooth CDF G can be represented

by $G(z) = \sum_{j=0}^{\infty} \pi_j^\star r_j(z)$ for any $z \in R$, where $\{\pi_j^\star\}_{j=0}^{\infty}$ is the unknown coefficients of the basis functions. In practice, to make our algorithm tractable, we truncate the sequence of the basis functions and only use the first $q+1$ basis functions for approximation, where q increases with sample size n at some rate. To approximate G, it then remains to provide an estimator for the unknown coefficients of the basis functions $\{\pi_j^\star\}_{j=0}^{q}$. Our estimation procedure for $\{\pi_j^\star\}_{j=0}^{q}$ shares similar intuition as the one that motivates the Nadaraya-Watson kernel estimator in the previous section. In particular, suppose for a moment that in the k-th round of update, we start with $\boldsymbol{\beta}_k$, which happens to be identical to the unknown true parameter $\boldsymbol{\beta}^\star$. In this case, define $\boldsymbol{r}_q(z) = (r_0(z), \cdots, r_q(z))^{\mathrm{T}}$ and $\boldsymbol{\pi}_q^\star = \left(\pi_1^\star, \cdots, \pi_q^\star\right)^{\mathrm{T}}$, we have that

$$y_i = G(z_{i,k}) + \varepsilon_i \approx \boldsymbol{r}_q^{\mathrm{T}}(z_{i,k})\, \boldsymbol{\pi}_q^\star + \varepsilon_i,$$

where recall that $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k$. The above relationship motivates the following OLS estimator for the sieve coefficients

$$\widehat{\boldsymbol{\pi}}_{q,n,k} = \left(\sum_{i=1}^{n} \boldsymbol{r}_q(z_{i,k})\, \boldsymbol{r}_q^{\mathrm{T}}(z_{i,k}) \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{r}_q(z_{i,k})\, y_i \right). \tag{1.15}$$

Given the estimator of the sieve coefficients $\widehat{\boldsymbol{\pi}}_{q,n,k}$, the unknown CDF G in the k-th round of update is approximated by

$$\widehat{G}(z|\boldsymbol{\beta}_k) = \boldsymbol{r}_q^{\mathrm{T}}(z)\, \widehat{\boldsymbol{\pi}}_{n,q,k}, \ -\infty < z < \infty. \tag{1.16}$$

Based on the estimated CDF $\widehat{G}(z|\boldsymbol{\beta}_k)$, the update of the parameter can be carried out based on (1.6). We iterate sequentially based on (1.15), (1.16) and (1.6) until some terminating conditions are satisfied. The resulting estimator is then labeled as the *sieve-based batch gradient descent estimator* (SBGD estimator). We summarize our algorithm as follows in algorithm 3.

Remark 1.7. *In the above SBGD procedure, we update the sieve parameter based on the OLS-type estimation. An alternative procedure can be based on the flexible Logit regression proposed by Hirano et al. (2003). The advantage of using flexible Logit regression is that the estimated CDF $\widehat{G}(z|\boldsymbol{\beta}_k)$ always falls between 0 and 1 for all z, which makes the update more stable. While the disadvantage of such update is that the flexible Logit regression is based on MLE, which does not allow for an analytical solution. Using numerical optimization to solve for the sieve coefficients in each round of update will add to additional computational burdens.*

Remark 1.8. *Compared with the KBGD algorithm, the SBGD procedure has at least two advantages.*

Algorithm 3: The SBGD Estimator

input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$, sequence of learning rate $\{\delta_k\}_{k=1}^\infty$, initial guess $\boldsymbol{\beta}_1$, the order of sieves q, sieve functions $\boldsymbol{r}(z) = r_0(z), \cdots, r_q(z)$, and terminating condition $\mathcal{T}$

output: The SBGD estimator $\widehat{\boldsymbol{\beta}}$

$k \leftarrow 1$;

while *The terminating condition $\mathcal{T}$ is not satisfied* **do**

  $\widehat{\boldsymbol{\pi}}_{q,n,k} \leftarrow \left(\sum_{i=1}^n \boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_i\right)$;

  for $i \leftarrow 1$ **to** n **do**

    $\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) \leftarrow \boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\widehat{\boldsymbol{\pi}}_{q,n,k}$;

  $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^n\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_{e,i}$;

  $k \leftarrow k + 1$;

$\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_k$;

On the one side, the sieve-based approximation for the unknown CDF is global and guarantees uniform approximation error rate. This allows us to update the parameter without performing any form of trimming as we did for the KBGD estimator. Moreover, this allows us to develop the asymptotic distribution of the SBGD estimator for the case of increasing dimensionality. On the otherhand, the KBGD procedure relies on the kernel estimation of CDF G at n data points, whose computational complexity of each update is of order $O\left(n^2\right)$. While the most time-consuming part of the SBGD procedure is the OLS procedure (1.15), whose computational complexity is of order $O\left(nq^2 + q^3\right)$. When $q/\sqrt{n} \to 0$, the computational burden of SBGD estimator will be substantially lower than that of KBGD estimator.

Define $R_q(z) = G(z) - \boldsymbol{r}^{\mathrm{T}}(z)\boldsymbol{\pi}_q^\star$, $\Gamma_{q,n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{r}_q\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $\Gamma_{q,n,k} = \Gamma_{q,n}(\boldsymbol{\beta}_k)$, and $\mathfrak{X}_{q,n}(z, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q,n}^{-1}(\boldsymbol{\beta})\boldsymbol{r}_q(z)\mathbf{X}_i\right)$. Through tedious algebra, we can show that the SBGD procedure has the following representation,

$$\begin{aligned}
\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k &- \frac{\delta_k}{n}\sum_{i=1}^n\left(\mathbf{X}_i - \mathfrak{X}_{q,n}(z_{i,k}, \boldsymbol{\beta}_k)\right)\left(G(z_{i,k}) - G(z_i^\star)\right) \\
&- \frac{\delta_k}{n}\sum_{i=1}^n \mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}(z_{i,k})\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{r}_q(z_{j,k})R_q(z_{j,k}) + \frac{1}{n}\sum_{i=1}^n \boldsymbol{r}_q(z_{j,k})\varepsilon_j\right) \\
&+ \frac{\delta_k}{n}\sum_{i=1}^n\left(R_q(z_{i,k})\mathbf{X}_i + \varepsilon_i\mathbf{X}_i\right), \qquad (1.17)
\end{aligned}$$

where recall that $z_i^\star = X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star$. To study the properties of the above procedure, we introduce some additional assumptions.

Assumption 1.6. *(i) There holds* $\max_{0\le j\le q}\|r_j\|_\infty \le D_{q,0}$, $\max_{0\le j\le q}\left\|r_j'\right\|_\infty \le D_{q,1}$, *and* $\max_{0\le j\le q}\left\|r_j''\right\|_\infty \le$

$D_{q,2}$; *(ii) Define* $\Gamma_q(\boldsymbol{\beta}) = \mathbb{E}\left(\boldsymbol{r}_q\left(X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$, *there hold* $\inf_{\boldsymbol{\beta}\in\mathcal{B}} \underline{\lambda}\left(\Gamma_q(\boldsymbol{\beta})\right) \geq \underline{\lambda}_\Gamma > 0$ *and* $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \overline{\lambda}\left(\Gamma_q(\boldsymbol{\beta})\right) \leq \overline{\lambda}_\Gamma < \infty$ *for all* q*; (iii) There hold* $\sup_{z\in R}\left|G(z) - \boldsymbol{r}^{\mathrm{T}}(z)\boldsymbol{\pi}_q^\star\right| \leq \mathcal{E}_{q,0}$ *and* $\sup_{z\in R}\left|G'(z) - \left(\boldsymbol{r}'(z)\right)^{\mathrm{T}}\boldsymbol{\pi}_q^\star\right| \leq \mathcal{E}_{q,1}$, *where* $\boldsymbol{r}'(z) = \left(r_0'(z), \cdots, r_q'(z)\right)^{\mathrm{T}}$.

For any $-\infty < z < \infty$, define the population counterpart of $\mathfrak{X}_{q,n}(z, \boldsymbol{\beta})$ as

$$\mathfrak{X}_q(z, \boldsymbol{\beta}) = \mathbb{E}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right)\right)\Gamma_q^{-1}(\boldsymbol{\beta})\,\boldsymbol{r}_q(z)\,\mathbf{X}\right).$$

Then we have the following lemma.

Lemma 1.4. *Define* $\chi_{1,n} = \sqrt{pq^2 D_{q,0}^4 \log\left(pqD_{q,0}D_{q,1}n\right)/n}$, *and* $\chi_{2,n} = \sqrt{p}qD_{q,0}^2\left(\chi_{1,n} + \mathcal{E}_{q,0}\right)$. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), and Assumption 1.6 hold, and moreover,* $\upsilon_G \geq 1$ *and the combination of* p, q *and* υ_G *guarantees that* $\chi_{1,n} \to 0$ *as* $n \to \infty$. *Then the following holds,*

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \delta_k \mathbb{E}\left[\left(\mathbf{X} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right)\right)\right)\right] + \delta_k \mathfrak{R}_{n,k},$$

where $\sup_{k\geq 1}\|\mathfrak{R}_{n,k}\| = O_p(\chi_{2,n})$.

Proof of Lemma 1.4. See subsection 1.7.1. □

Obviously, Lemma 1.4 provides a parallel result to (1.10). In particular, define

$$\Psi_q(t, \boldsymbol{\beta}) = \mathbb{E}\left[G'\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right) + t\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}^{\mathrm{T}}\right)\right],$$

under all the conditions imposed in Lemma 1.4, we have that

$$\Delta\boldsymbol{\beta}_{k+1} = \left\{\int_0^1 \left(I_p - \delta_k \Psi_q\left(t, \boldsymbol{\beta}_k\right)\right) dt\right\}\Delta\boldsymbol{\beta}_k + \delta_k \mathfrak{R}_{n,k}. \tag{1.18}$$

Obviously, (1.18) is also a parallel result to (1.11). As a result, to ensure that (1.18) actually constitutes a contraction for $\|\Delta\boldsymbol{\beta}_k\|$, we impose the following assumption that is similar to Assumption 2.5.

Assumption 1.7. *For any* $q \geq 0$, *there hold*

$$\inf_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}} \underline{\lambda}\left(\Psi_q(t, \boldsymbol{\beta}) + \Psi_q^{\mathrm{T}}(t, \boldsymbol{\beta})\right) \geq \underline{\lambda}_\Psi > 0,$$

$$\sup_{0\leq t\leq 1,\boldsymbol{\beta}\in\mathcal{B}}\underline{\lambda}\left(\Psi_q\left(t,\boldsymbol{\beta}\right)+\Psi_q^{\mathrm{T}}\left(t,\boldsymbol{\beta}\right)\right)\geq\overline{\lambda}_{\Psi}<\infty.$$

Based on the above assumptions, we have the following result.

Theorem 1.7. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), Assumption 1.6 and Assumption 1.7 hold, $\upsilon_G\geq 1$, and the combination of p, q and υ_G guarantees that $\chi_{1,n}\to 0$ as $n\to\infty$. Suppose moreover that the learning rate is chosen such that $\delta_k=\delta$ with $0<\delta<\min\left\{1/\left(2\underline{\lambda}_{\Psi}\right),\underline{\lambda}_{\Psi}/\left(2\left\|G'\right\|_{\infty}^2p^2\left\{1+\underline{\lambda}_{\Gamma}^{-1}qD_{q,0}^2\right\}^2\right)\right\}$, and that $\boldsymbol{\beta}$ is updated based on algorithm 3. Define*

$$k_{1,n}^{SBGD}=\frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right)-\log\left(\chi_{2,n}\right)}{-\log\left(1-\underline{\lambda}_{\Psi}\delta/4\right)},$$

then we have that

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\|=O_p\left(\chi_{2,n}\right).$$

Proof of Theorem 1.7. See subsection 1.7.2. □

According to Theorem 1.7, when $\chi_{2,n}\to 0$ as $n\to\infty$, the SBGD estimator is consistent as long as the number of updates exceeds $k_{1,n}^{SBGD}$. Based on such consistent estimator, we are ready to establish the asymptotic normality of our SBGD estimator. Apply the mean value theorem to (1.17), we have that

$$\begin{aligned}\Delta\boldsymbol{\beta}_{k+1}=&\left\{I_p-\delta_k\int_0^1\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^{\star}+t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_k\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathfrak{X}_{q,n}\left(z_{i,k},\boldsymbol{\beta}_k\right)\mathbf{X}_i^{\mathrm{T}}\right)dt\right\}\Delta\boldsymbol{\beta}_k\\&-\frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)R_q\left(z_{j,k}\right)+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)\varepsilon_j\right)\\&+\frac{\delta_k}{n}\sum_{i=1}^{n}\left(R_q\left(z_{i,k}\right)\mathbf{X}_i+\varepsilon_i\mathbf{X}_i\right).\end{aligned}$$

Define $\Psi_q^{\star}=\mathbb{E}\left[G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^{\star}\right)\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathfrak{X}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^{\star}\right),\boldsymbol{\beta}^{\star}\right)\mathbf{X}^{\mathrm{T}}\right)\right]$ and $\mathfrak{V}_q=\mathbb{E}\left(\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(z_i^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)$. Similar to Lemma 1.2 and Lemma 1.3, we provide two additional lemmas that are useful to understand the above algorithm.

Lemma 1.5. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), and Assumption 1.6 hold, $\upsilon_G\geq 2$ and the combination of p, q and υ_G guarantees that $\chi_{1,n}\to 0$ as $n\to\infty$. Then for any*

sequence $\{\mathcal{B}_n\}_{n=1}^{\infty}$ *with* $\mathcal{B}_n \subseteq \mathcal{B}$ *we have that*

$$\sup_{0\leq t\leq 1,\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^{\star}+t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)-\Psi_q^{\star}\right\|$$
$$=O_p\left(pqD_{q,0}^2\chi_{1,n}+\sqrt{p^3}q^2D_{q,0}^3D_{q,1}\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\|\Delta\boldsymbol{\beta}\|\right).$$

Proof of Lemma 1.5. See subsection 1.7.1. □

Lemma 1.6. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), Assumption 1.6, and Assumption 1.7 hold, and the combination of* p, q *and* υ_G *guarantees that* $\chi_{1,n}\to 0$ *as* $n\to\infty$. *Define* $\boldsymbol{r}_{q,i,k}=\boldsymbol{r}_q\left(z_{i,k}\right)$, *and* $R_{q,i,k}=R_q\left(z_{i,k}\right)$. *Also define*

$$\chi_{3,n}=\sqrt{p^2qD_{q,1}^2\log\left(pqD_{q,2}n\right)/n},$$

then we have that

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}R_{q,j,k}+\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j\right)+\right.$$
$$\left.\frac{1}{n}\sum_{i=1}^{n}R_q\left(z_{i,k}\right)\mathbf{X}_i-\frac{1}{n}\sum_{i=1}^{n}\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\varepsilon_j\right\|=O_p\left(\chi_{4,n}\right),$$

where $\chi_{4,n}=\sqrt{p}qD_{q,0}^2\mathcal{E}_{q,0}+\sqrt{pq}D_{q,0}\chi_{2,n}\chi_{3,n}+\chi_{2,n}\sqrt{p^2q^4D_{q,0}^6D_{q,1}^2\left(\log q\right)/n}$.

Proof of Lemma 1.6. See subsection 1.7.1. □

Based on the above two lemmas, we are now ready to study the asymptotic distribution of the SBGD estimator.

Theorem 1.8. *Suppose that Assumption 1.1, Assumption 1.2(i)-(iii), Assumption 1.6 and Assumption 1.7 hold,* $\upsilon_G\geq 2$, *the combination of* p, q *and* υ_G *guarantees that* $\chi_{1,n}\to 0$ *as* $n\to\infty$, *and that* $\boldsymbol{\beta}$ *is updated based on algorithm 3. We have that*

(i) There holds

$$\Delta\boldsymbol{\beta}_{k+1}=\left(I_p-\delta\Psi_q^{\star}\right)\Delta\boldsymbol{\beta}_k+\frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\right)\varepsilon_i+\widetilde{\mathfrak{R}}_{n,k},$$

where $\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\widetilde{\mathfrak{R}}_{n,k}\right\|=O_p\left(\chi_{5,n}\right)$ *with*

$$\chi_{5,n}=\sqrt{p}qD_{q,0}^2\left(p+qD_{q,0}D_{q,1}\right)\chi_{2,n}^2+\chi_{4,n};$$

(ii) Define $\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}_{k+k_{1,n}^{SBGD}+k_{2,n}^{SBGD}+1}$ *with*

$$k_{2,n}^{SBGD}=\frac{-\log\chi_{2,n}+\log\sqrt{n}}{-\log\left(1-\underline{\lambda}_{\Psi}\delta/4\right)},$$

and any $k\geq 1$. *If the combination of* p, q *and* v_G *further guarantees that* $\sqrt{n}\chi_{5,n}\to 0$ *as* $n\to\infty$, *we have that*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\star}\right)=\Psi_q^{\star-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\right)\varepsilon_i+o_p\left(n^{-\frac{1}{2}}\right).$$

Then for any $p\times 1$ *vector* ρ *such that* $\|\rho\|<\infty$ *and* $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho^{\mathrm{T}}\Psi_q^{\star-1}\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\right)\varepsilon_i\to_d N\left(0,\sigma_S^2\left(\rho\right)\right)$ *with*

$$\sigma_S^2\left(\rho\right)=\lim_{n\to\infty}\rho^{\mathrm{T}}\Psi_q^{\star-1}\mathbb{E}\left\{G\left(z_i^{\star}\right)\left(1-G\left(z_i^{\star}\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}\right\}\left(\Psi_q^{\star-1}\right)^{\mathrm{T}}\rho,$$

there holds

$$\sqrt{n}\rho^{\mathrm{T}}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\star}\right)\to_d N\left(0,\sigma_S^2\left(\rho\right)\right)$$

.

Proof of Theorem 1.8. See subsection 1.7.2. □

We now provide the estimator for the variance.

Theorem 1.9. *Suppose that all the conditions listed in Theorem 1.8 hold and* $pq^2D_{q,0}^4\mathcal{E}_{q,1}\to 0$ *as* $n\to 0$. *Let* $\widehat{\boldsymbol{\beta}}$ *be as defined as in Theorem 1.8. Define* $\widehat{\boldsymbol{r}}_{q,i}=\boldsymbol{r}_q\left(z\left(\mathbf{X}_{e,i},\widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{r}}_{q,i}'=\boldsymbol{r}_q'\left(z\left(\mathbf{X}_{e,i},\widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{\pi}}_q=\left(\sum_{i=1}^{n}\widehat{\boldsymbol{r}}_{q,i}\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\boldsymbol{r}}_{q,i}y_i\right)$, $\widehat{G}_i=\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\widehat{\boldsymbol{\pi}}$, $\widehat{G}_i'=\widehat{\boldsymbol{r}}_{q,i}'^{\mathrm{T}}\widehat{\boldsymbol{\pi}}_q$, $\widehat{\Psi}_{q,i}^{\star}=\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i'\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathfrak{X}_{q,n}\left(\widehat{z}_i,\widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i^{\mathrm{T}}\right)$, $\widehat{\mathfrak{X}}_{q,i}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j\widehat{\boldsymbol{r}}_{q,j}^{\mathrm{T}}\Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{r}}_{q,i}$, *and*

$$\widehat{\sigma}_S^2\left(\rho\right)=\rho^{\mathrm{T}}\widehat{\Psi}_q^{\star-1}\frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i-\widehat{\mathfrak{X}}_{q,i}\right)\left(\mathbf{X}_i-\widehat{\mathfrak{X}}_{q,i}\right)^{\mathrm{T}}\right\}\left(\widehat{\Psi}_q^{\star-1}\right)^{\mathrm{T}}\rho,$$

Then for any $p \times 1$ vector ρ such that $\|\rho\| < \infty$, there holds

$$\left|\widehat{\sigma}_S^2(\rho) - \sigma_S^2(\rho)\right| \to_p 0.$$

Proof of Theorem 1.9. See subsection 1.7.2. □

We finally provide some remarks on the empirical applications of the SBGD estimator.

Remark 1.9. *For the choice of sieve functions, we can use polynomial series for the case where the error term u_i has bounded support and Hermite polynomials for the case where u_i has unbounded support. Note that when using polynomial series $\left\{1, z, z^2, \cdots, z^q\right\}$, the correlation between the sieve functions increases as the approximation order q increases, which may lead to a violation of Assumption 1.6(ii). To improve the finite sample performance of our method, we recommend using Chebyshev or Legendre polynomials. Moreover, in the case where u_i has unbounded support, following Bierens (2014), we recommend first conducting the following transformation $G(z) = \widetilde{G}(T(z))$, where $T : R \mapsto [-1, 1]$ is a differentiable function, and then using standard Chebyshev or Legendre polynomials to approximate $\widetilde{G}$. For example, in our following simulations and empirical applications in Section 3.6, we use $T(z) = 2\pi^{-1}\arctan(z)$. For the uniform error bound of truncated Legendre polynomials, see Wang and Xiang (2012).*

1.4 Monte Carlo Experiments

This section conducts Monte Carlo simulations to study the performance of our KBGD and SBGD estimators. We focus on two aspects of our estimators. First we study the finite-sample properties of the KBGD estimator, including the bias and the root mean squared error (RMSE). Let the j-th argument of the true parameter be $\beta_j^\star$, and the simulation is repeated R times, where its estimator in the r-th round of simulation is $\widehat{\beta}_j^r$, then the bias and RMSE are respectively given by Bias $= |\frac{1}{R}\sum_{r=1}^R(\widehat{\beta}_j^r - \beta_j^\star)|$ and RMSE $= \sqrt{\sum_{r=1}^R(\widehat{\beta}_j^r - \beta_j^\star)^2/R}$. We also investigate whether the confidence interval based on the asymptotic distribution has good coverage rate. We consider nominal coverage rate $\alpha = 0.95$, so the confidence interval for $\beta_j^\star$ in the r-th round of repetition is given by $CI_j^r = [\widehat{\beta}_j^r - 1.96 \cdot \widehat{\text{std}}_j^r, \widehat{\beta}_j^r + 1.96 \cdot \widehat{\text{std}}_j^r]$, where $\widehat{\text{std}}_j^r$ is the estimated standard deviation of $\widehat{\beta}_j^r$. The actual coverage rate is then given by $CR = \frac{1}{R}\sum_{r=1}^R I(\beta_j^\star \in CI_j^r)$.

We are also interested in how sensitive our estimators are to the initial guess of the true parameter.

Table 1.1: Finite Sample Performance of KBGD and SBGD Estimators

	Bias		RMSE		CR		Bias		RMSE		CR	
	KBGD	SBGD	KBGD	SBGD	KBGD	SBGD	KBGD	SBGD	KBGD	SBGD	KBGD	SBGD
	$n = 2500$						$n = 5000$					
β_1	0.0024	0.0031	0.1193	0.1240	0.9600	0.9680	0.0047	0.0005	0.0844	0.0867	0.9500	0.9600
β_2	0.0002	0.0055	0.1255	0.1336	0.9480	0.9500	0.0031	0.0074	0.0846	0.0878	0.9520	0.9540
β_3	0.0136	0.0260	0.1544	0.1791	0.9480	0.9460	0.0004	0.0074	0.1053	0.1112	0.9320	0.9320
β_4	0.0093	0.0213	0.1551	0.1706	0.9500	0.9440	0.0012	0.0095	0.1035	0.1117	0.9600	0.9500
β_5	0.0257	0.0482	0.2511	0.2968	0.9540	0.9400	0.0007	0.0168	0.1648	0.1889	0.9400	0.9480
β_6	0.0236	0.0477	0.2502	0.2860	0.9480	0.9580	0.0121	0.0269	0.1723	0.1931	0.9540	0.9360
β_7	0.0500	0.0964	0.4513	0.5416	0.9640	0.9420	0.0051	0.0352	0.3083	0.3525	0.9440	0.9420
β_8	0.0447	0.0920	0.4662	0.5441	0.9360	0.9520	0.0098	0.0394	0.3121	0.3477	0.9420	0.9440
β_9	0.0242	0.0454	0.2921	0.3303	0.9480	0.9500	0.0072	0.0048	0.1840	0.1909	0.9540	0.9560
β_{10}	0.0168	0.0338	0.1881	0.2223	0.9520	0.9440	0.0030	0.0147	0.1247	0.1402	0.9440	0.9380

NOTE: For KBGD estimator, we use fourth-order Epanechinikov kernel to construct the Nadaraya-Watson estimator. We choose $\delta = 1$. In each round of iteration, the bandwidth h_n is chosen as $h_n = \sigma_{\widehat{z}} \cdot n^{-1/5}$, where n is sample size, $\sigma_{\widehat{z}}$ is the standard deviation of $z_{i,k}$, and $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k$. For SBGD estimator, we choose $q = 9$ and use Legendre polynomials with transformation discussed in Remark 1.9. For both estimators, the stopping rule is either $\max_{1 \le j \le p} |\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}| < 10^{-5}$ or $k \ge 20000$. The above also applies to our empirical analysis in Section 3.6. Trimming is ignored during all the simulations. Due to the outliers of the simulation, we trim out the lower and upper 2% simulation results and calculate the bias and RMSE.

In each repetition of our simulation, we consider three different initial guesses: the true parameter vector, the parameter vector estimated based on the Logit regression, and the parameter with all elements being zeros. If the estimation results starting from different initial guesses are close or even identical to each other, the estimation methods are insensitive to the initial guesses and thus are robust in terms of computation. Denote $\widehat{\boldsymbol{\beta}}_T^r$, $\widehat{\boldsymbol{\beta}}_L^r$, and $\widehat{\boldsymbol{\beta}}_Z^r$ as the estimators with starting points being true parameter, Logit estimator, and vector of zeros. We use $S_L = \sqrt{\frac{1}{R}\sum_{i=1}^n ||\widehat{\boldsymbol{\beta}}_L^r - \widehat{\boldsymbol{\beta}}_T^r||^2}$ and $S_Z = \sqrt{\frac{1}{R}\sum_{i=1}^n ||\widehat{\boldsymbol{\beta}}_Z^r - \widehat{\boldsymbol{\beta}}_T^r||^2}$ as the measurement of the sensitivity. To compare the performance of our method with the existing estimators, we also consider Ichimura's semiparametric least squares (SLS) estimator (Ichimura, 1993) and Klein and Spady's semiparametric maximum likelihood (SMLE) estimator (Klein and Spady, 1993).

We consider data generating process $y_i = I(X_{0,i} + \beta_1^\star X_{1,i} \cdots + \beta_{10}^\star X_{10,i} - u_i > 0), i = 1, 2, \cdots, n$, where data are i.i.d over i, and $X_{0,i}, X_{1,i}, \cdots, X_{10,i}, u_i$ are also independent. We set

$$\boldsymbol{\beta}^\star = (1, 0.5, -0.5, 1, -1, 2, -2, 4, -4, 1.5, -1.5)^{\mathrm{T}},$$

$X_{j,i} \sim N(0,1)$ for $0 \le j \le 8$, $X_{9,i} \sim$ Bernoulli $(1/2)$, $X_{10,i} \sim$ Poisson (2), and $u_i \sim Cauchy$. We consider two sample sizes $n = 2500$ and 5000. Finally, for finite-sample performance, we repeat the simulation 500 times; for sensitivity analysis, we repeat 100 times.

Table 1.1 reports the finite-sample properties of our estimators. It can be seen that our estimators

Table 1.2: Sensitivity of KBGD and SBGD Estimators: Fixed Coefficients

		Sensitivity		Running Time		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0242	0.0198	113.21	79.120	158.91
	SBGD	0.0175	0.0259	0.9504	0.9482	1.1587
	SLS	0.8732	251.58	35.695	37.210	35.104
	SMLE	0.9362	318.41	34.515	33.704	31.078
$n = 5000$	KBGD	0.0241	0.0175	157.48	87.954	230.07
	SBGD	0.0189	0.0282	1.4644	1.4722	1.9074
	SLS	0.6870	871.58	46.402	44.647	41.486
	SMLE	0.7343	507.69	44.563	43.256	35.904

NOTE: SLS refers to semiparametric least squares estimator, and SMLE refers to semiparametric maximum likelihood estimator. The running time is all in seconds. Due to the outliers of the simulation, we trim out the lower and upper 2% simulation results and calculate the corresponding results. The above also applies to Table 3.3.

Table 1.3: Sensitivity of KBGD and SBGD Estimators: Random Coefficients

		Sensitivity		Running Time		
	Method	S_L	S_Z	True	Logit	Zeros
$n = 2500$	KBGD	0.0270	0.0214	122.00	74.433	166.94
	SBGD	0.0123	0.0246	1.0132	0.8252	1.2044
	SLS	0.9178	500.24	34.864	35.571	34.065
	SMLE	0.9956	533.58	34.334	32.520	29.473
$n = 5000$	KBGD	0.0234	0.0232	163.74	91.449	247.49
	SBGD	0.0077	0.0234	1.5529	1.4377	1.9217
	SLS	0.6796	10737	43.935	41.420	46.449
	SMLE	0.6821	698.63	43.616	44.825	37.763

work well in finite sample cases. Both estimators have small bias, whose RMSE decrease with the increase of sample size. Moreover, the confidence interval constructed based on the asymptotic variance and normal approximation has actual coverage rate that is quite close to the nominal rate 0.95.

Table 1.2 reports the sensitivity of our estimators to the starting points. We can see that for both KBGD and SBGD estimators, S_L and S_Z are close to zero, indicating that the resulting estimators starting from Logit estimator or zeros are almost identical to the ones starting from the unknown true parameter. Such a result demonstrates that our algorithms are robust to different initial guesses. On the contrary, the SLS and SMLE are both sensitive to the initial guess. As we can see, the estimators starting from parametric Logit regression differ significantly from those starting from the unknown true parameter, and such difference even explodes when we consider estimators starting from the origin point. The above results highlight the numerical robustness of our estimators.

The robustness of our algorithm might also be sensitive to the setups of coefficients. To check whether this is the case, instead of using the fixed parameters specified before, in each round of simulation we randomly draw true parameter $\boldsymbol{\beta}^\star$ as follows $\beta_1^\star, \beta_2^\star, \beta_9^\star, \beta_{10}^\star \sim N(0,1)$, $\beta_3^\star, \beta_4^\star, \beta_5^\star, \beta_6^\star \sim 2N(0,1)$,

and $\beta_7^\star, \beta_8^\star \sim 4N(0,1)$. The simulation results are reported in Table 1.3. We can see that the results are similar to those under fixed parameters, indicating that our algorithm is robust to initial point under different parameter setups.

1.5 Empirical Application

As an empirical illustration of our new methods, this section applies our KBGD and SBGD estimation procedures to study how education affects the risk aversion. In the existing researches, it's extensively documented that, on the individual level, risk aversion is significantly correlated with the level of education, although the directions of correlation are mixed, see Outreville (2015) for a comprehensive review. In this study, we investigate how educational background of the family affects the risk aversion of the household as well as household-level investing behaviors. We use the national survey data from 2019 China Household Financial Survey Project (CHFS) (Gan et al., 2014), which provides household-level information over demographics, asset and debt, income and consumption, social security and insurance, and various household's subjective preferences. The dependent variable we are interested in is the degree of risk aversion of the household. In particular, y_i is constructed to take value of 0 if the i-th household is completely against any form of risks and thus is described as being extremely risk averse; it takes value of 1 if the family is willing to bear some form of risks when making investments. We study how the probability of $y_i = 1$ is affected by a set of factors based on the binary choice model. The key factor that we are particularly interested in is the educational backgrounds, which is defined as the year of education of the head of the household. We also consider a set of other control variables including gender, ethnicity, health conditions, marital status, region of residence, economic knowledge, total income and total asset, whose impacts on the risk aversion are of interest on their own right. See Yao (2023) for detailed discussion on the construction of the data sets.

Before estimation, we normalize all the continuous variables so that the resulting variables all have zero mean and unity variance. To provide a comparison to the semiparametric estimation results, we first conduct parametric Logit regression and report the normalized coefficients in regression (I) in Table 1.4. We then conduct KBGD and SBGD estimation and report the estimated coefficients of education in (II) and (III). As we can see from Table 1.4, no matter which estimation methods we use, the coefficient of educational background is estimated to be positive with significance at 1% level. This implies that, holding other conditions fixed, on average an increase in the year of

Table 1.4: Estimation Results

	(I)	(II)	(III)
Estd. Coefficients	2.5543***	2.4832***	2.4647***
	(0.1070)	(0.3638)	(0.3239)
Num. of Obs.	26906	26906	26906
Estimation Methods	Logit	KBGD	SBGD
Running Time	1.4276	8573.1	40.9941
Num. of Iteration	–	14996	12986

Note: For Logit regression, we report the coefficient of education divided by that of total asset. For semiparametric estimation, we normalize the coefficient of total asset to be 1. The standard deviations are reported in the brackets below the coefficients. *** indicates significance at 1% level. For both KBGD and SBGD estimators, we choose $\delta_k = 1$. For KBGD estimator, we choose $h_n = C \cdot n^{-1/5}$ with $C = C_k = \text{std}(z_{i,k})$, and use the fourth-order Epanechinikov kernel. For SBGD estimator, we choose $q = 9$ and use Legendre polynomials with transformation discussed in Remark 1.9. The starting point of iteration for both KBGD and SBGD estimators is chosen as the origin point with all arguments being 0. The stopping rule is set as $\max_{1\leq j\leq p} |\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}| < \varrho$ with $\varrho = 10^{-5}$. Finally, the running time is in second.

education of the head in the households leads to the increase of willingness to bear risks. Comparing the semiparametric estimation results with that of Logit regression, we can see that the KBGD and SBGD estimators are close to each other, which are both smaller than that of Logit regression, indicating that parametric estimation might suffer from model misspecification and lead to an over-estimation of the impacts of education on risk aversion. We finally compare the computation time of each method. We can see that both KBGD and SBGD estimators take much longer to converge compared with the parametric estimation. Comparatively, the SBGD algorithm is significantly faster than the KBGD algorithm, which takes over two hours to converge. This result supports the use of SBGD algorithm when there are data of large scale.

1.6 Conclusions

In this paper, we proposed new estimation procedures for binary choice and monotonic index models with increasing dimensions. Existing semiparametric estimation procedures for this model cannot be implemented in practice when the number of regressors is large. In contrast, our algorithmic based procedures can be used for many regressor models as it involves convex optimization at each iteration of the procedure. We show this iterative procedure also has desirable asymptotic properties when the number of regressors increases with the sample size in ways that are standard in big data literature.

1.7 Technical Details

1.7.1 Lemmas and Proofs

This part provides some lemmas that will be used during the establishment of our results in the main context. If not otherwise stated, the dimension p of covariate $\mathbf{X}$ is allowed to increase with sample size n.

Lemma 1.7. *Consider i.i.d. random variables* $\{U_i\}_{i=1}^n$ *on probability space* $(\Omega, \mathscr{A}, P)$ *and* $d_1 \times d_2$ *matrix* $A(U,\theta) : \Omega \times \Theta \to R^{d_1 \times d_2}$ *with* $\Theta \subseteq R^p$ *being compact,* $\sup_{U\in\Omega,\theta\in\Theta} \|A_{s,t}(U,\theta)\| \le D_{A,0}$ *and* $\sup_{U\in\Omega} \|A_{s,t}(U,\theta_1) - A_{s,t}(U,\theta_2)\| \le D_{A,1}\|\theta_1 - \theta_2\|$ *uniformly for all* $1 \le s \le d_1$ *and* $1 \le t \le d_2$. *Then there holds*

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta) - \mathbb{E}A(U_i,\theta) \right\| = O_p\left(\sqrt{\frac{pd_1d_2D_{A,0}^2 \log(d_1d_2D_{A,1}n)}{n}} \right).$$

Proof of Lemma 1.7. Note that

$$\begin{aligned}
\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta) - \mathbb{E}A(U_i,\theta) \right\| \le & \max_{1\le b\le B} \left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta_b) - \mathbb{E}A(U_i,\theta_b) \right\| \\
& + \max_{1\le b\le B} \sup_{\|\theta-\theta_b\|\le \frac{C}{\sqrt[p]{B}}} \left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta) - \frac{1}{n}\sum_{i=1}^n A(U_i,\theta_b) \right\| \\
& + \max_{1\le b\le B} \sup_{\|\theta-\theta_b\|\le \frac{C}{\sqrt[p]{B}}} \|\mathbb{E}A(U_i,\theta) - \mathbb{E}A(U_i,\theta_b)\|.
\end{aligned}$$

For the first term, we have that

$$\begin{aligned}
& P\left(\max_{1\le b\le B} \left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta_b) - \mathbb{E}A(U_i,\theta_b) \right\| > \tau \right) \\
& \le \sum_{b=1}^B P\left(\left\| \frac{1}{n}\sum_{i=1}^n A(U_i,\theta_b) - \mathbb{E}A(U_i,\theta_b) \right\| > \tau \right) \\
& \le \sum_{b=1}^B P\left(\max_{1\le s\le d_1} \max_{1\le t\le d_2} \left| \frac{1}{n}\sum_{i=1}^n A_{s,t}(U_i,\theta_b) - \mathbb{E}A_{s,t}(U_i,\theta_b) \right| > \frac{\tau}{\sqrt{d_1d_2}} \right) \\
& \le \sum_{b=1}^B \sum_{s=1}^{d_1}\sum_{t=1}^{d_2} P\left(\left| \frac{1}{n}\sum_{i=1}^n A_{s,t}(U_i,\theta_b) - \mathbb{E}A_{s,t}(U_i,\theta_b) \right| > \frac{\tau}{\sqrt{d_1d_2}} \right) \\
& \le \sum_{b=1}^B \sum_{s=1}^{d_1}\sum_{t=1}^{d_2} 2\exp\left(-Cn\tau^2/\left(d_1d_2D_{A,0}^2\right)\right) = 2\exp\left(C\log(Bd_1d_2) - Cn\tau^2/\left(d_1d_2D_{A,0}^2\right)\right),
\end{aligned}$$

indicating that

$$\max_{1\leq b\leq B}\left\|\frac{1}{n}\sum_{i=1}^{n}A\left(U_i,\theta_b\right)-\mathbb{E}A\left(U_i,\theta_b\right)\right\|=O_p\left(\sqrt{\frac{d_1d_2D_{A,0}^2\log\left(Bd_1d_2\right)}{n}}\right).$$

On the other side, for the second term we have that

$$\begin{aligned}&\max_{1\leq b\leq B}\sup_{\|\theta-\theta_b\|\leq\frac{C}{\sqrt[p]{B}}}\left\|\frac{1}{n}\sum_{i=1}^{n}A\left(U_i,\theta\right)-\frac{1}{n}\sum_{i=1}^{n}A\left(U_i,\theta_b\right)\right\|\\&\leq\sqrt{d_1d_2}\max_{1\leq s\leq d_1}\max_{1\leq t\leq d_2}\sup_{U\in\Omega}\sup_{\|\theta-\theta_b\|\leq\frac{C}{\sqrt[p]{B}}}\left|A_{s,t}\left(U,\theta\right)-A_{s,t}\left(U,\theta_b\right)\right|\leq\frac{\sqrt{d_1d_2}D_{A,1}}{\sqrt[p]{B}}.\end{aligned}$$

The same bound holds for the third term. Then let $B=(\sqrt{n}D_{A,1})^p$, we finish the proof. □

Lemma 1.8. *If Assumption 1.1, Assumption 1.2(i)-(iii), and Assumption 1.4 hold with* $\min\{\upsilon_G,\upsilon_f\}\geq 2$*, then there exists a constant* C *that does not depend on* $\mathbf{X},z,\boldsymbol{\beta}$ *such that the following hold*

(i) $\sup_{\mathbf{X},z,\boldsymbol{\beta}}\left|\partial^s f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)/\partial z^s\right|\leq C$ *for* $0\leq s\leq\upsilon_f$*;*

(ii) $\sup_{z,\boldsymbol{\beta}}\left|\partial^s f_{z}\left(z|\,\boldsymbol{\beta}\right)/\partial z^s\right|\leq C$ *for* $0\leq s\leq\upsilon_f$*;*

(iii) $\sup_{\mathbf{X},z,\boldsymbol{\beta}}\left\|\partial f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|\leq C\sqrt{p}$*;*

(iv) $\sup_{\mathbf{X},z,\boldsymbol{\beta}}\left\|\partial^2 f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\right\|\leq Cp$*;*

(v) $\left\|\partial f_{z}\left(z|\,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|\leq C\sqrt{p}$*;*

(vi) $\left\|\partial^2 f_{z}\left(z|\,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\right\|\leq Cp$*;*

(vii) $\sup_{z,\boldsymbol{\beta},f_z(z|\boldsymbol{\beta})\neq 0}\left|\partial^s L\left(z,\boldsymbol{\beta}\right)/\partial z^s\right|\leq C$ *for* $0\leq s\leq\min\{\upsilon_G,\upsilon_f\}$*;*

(viii) $\sup_{z,\boldsymbol{\beta},f_z(z|\boldsymbol{\beta})\neq 0}\left\|\partial L\left(z,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|\leq C\sqrt{p}$;

(ix) $\sup_{z,\boldsymbol{\beta},f_z(z|\boldsymbol{\beta})\neq 0}\left\|\partial^2 L\left(z,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}\right\|\leq Cp$*;*

(x) $\sup_{\mathbf{X}_e,\boldsymbol{\beta},f_z(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta})\neq 0}\int_{\mathcal{X}}\left\|\partial W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\widetilde{\mathbf{X}}\leq C\sqrt{p}$.

Proof. To prove Lemma 1.8(i) and Lemma 1.8(ii), we note that for any $0\leq s\leq\upsilon_f$,

$$\frac{\partial^s f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)}{\partial z^s}=\left.\frac{\partial^s f_e\left(X_0,\mathbf{X}\right)}{\partial X_0^s}\right|_{X_0=z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}},$$

and

$$\frac{\partial^s f_z(z|\boldsymbol{\beta})}{\partial z^s} = \int_{\mathcal{X}} \left[\frac{\partial^s f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})}{\partial X_0^s} \right] d\mathbf{X}.$$

Since $f_e(\mathbf{X}_e)$ has up to υ_f-th bounded derivatives over $\mathcal{X}_e$ according to Assumption 2.4(ii) and X_j is bounded by 1 for all $1 \leq j \leq p$ according to Assumption 2.2(i), Lemma 1.8(i) and Lemma 1.8(ii) hold.

Similarly, note that

$$\frac{\partial f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\left[\frac{\partial f_e(X_0, \mathbf{X})}{\partial X_0}\Big|_{X_0 = z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} \right] \mathbf{X},$$

$$\frac{\partial^2 f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} = \left[\frac{\partial^2 f_e(X_0, \mathbf{X})}{\partial X_0^2}\Big|_{X_0 = z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} \right] \mathbf{X}\mathbf{X}^{\mathrm{T}},$$

$$\frac{\partial f_z(z|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\int_{\mathcal{X}} \left[\frac{\partial f_e(X_0, \mathbf{X})}{\partial X_0}\Big|_{X_0 = z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} \right] \mathbf{X} d\mathbf{X},$$

$$\frac{\partial^2 f_z(z|\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathrm{T}}} = \int_{\mathcal{X}} \left[\frac{\partial^2 f_e(X_0, \mathbf{X})}{\partial X_0^2}\Big|_{X_0 = z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} \right] \mathbf{X}\mathbf{X}^{\mathrm{T}} d\mathbf{X},$$

we validate Lemma 1.8(iii)-Lemma 1.8(vi).

To prove Lemma 1.8(vii), note that

$$\begin{aligned}
\left| \frac{\partial^s L(z, \boldsymbol{\beta})}{\partial z^s} \right| &\leq C \sum_{j=0}^{s} \left| \int_{\mathcal{X}} G^{(j)}\left(z - \mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta} \right) \frac{\partial^{s-j} f_{\mathbf{X}|z}(\mathbf{X}|\, z, \boldsymbol{\beta})}{\partial z^{s-j}} d\mathbf{X} \right| \\
&\leq C \sum_{j=0}^{s} \left\| G^{(j)} \right\|_{\infty} \cdot \left(\int_{\mathcal{X}} \left| \frac{\partial^{s-j} f_{\mathbf{X}|z}(\mathbf{X}|\, z, \boldsymbol{\beta})}{\partial z^{s-j}} \right| d\mathbf{X} \right).
\end{aligned}$$

According to Assumption 2.2(iii), $\left\| G^{(j)} \right\|_{\infty}$ is bounded for all $0 \leq j \leq \upsilon_G$. Then it remains to show that $\int_{\mathcal{X}} \left| \partial^{s-j} f_{\mathbf{X}|z} / \partial z_{\infty}^{s-j} \right| d\mathbf{X}$ is also upper bounded for all $0 \leq j \leq \upsilon_f$. When $j = s$, we have that $\int_{\mathcal{X}} \left| \partial^{s-j} f_{\mathbf{X}|z}(\mathbf{X}|\, z, \boldsymbol{\beta}) / \partial z^{s-j} \right| d\mathbf{X} = 1$. When $j = s-1$, define $\mathbb{X}(z, \boldsymbol{\beta}) = \left\{ \mathbf{X} : \left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X} \right) \in \mathcal{X}_e \right\}$. We have that

$$\begin{aligned}
&\int_{\mathcal{X}} \left| \frac{\partial f_{\mathbf{X}|z}(\mathbf{X}|\, z, \boldsymbol{\beta})}{\partial z} \right| d\mathbf{X} \\
&= \int_{\mathcal{X}} \left| \frac{\partial f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta}) / \partial z}{\int_{\mathcal{X}} f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})\, d\mathbf{X}} - \frac{f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta}) \int_{\mathcal{X}} \left(\partial f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta}) / \partial z \right) d\mathbf{X}}{\left(\int_{\mathcal{X}} f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})\, d\mathbf{X} \right)^2} \right| d\mathbf{X} \\
&\leq \frac{2 \int_{\mathcal{X}} \left| \partial f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta}) / \partial z \right| d\mathbf{X}}{\int_{\mathcal{X}} f_{\mathbf{X},z}(\mathbf{X}, z|\boldsymbol{\beta})\, d\mathbf{X}} \leq \frac{2 \left\| \partial f_{\mathbf{X},z} / \partial z \right\|_{\infty} m\left(\mathbb{X}(z, \boldsymbol{\beta}) \right)}{\zeta^{-1} m\left(\mathbb{X}(z, \boldsymbol{\beta}) \right)} \leq C
\end{aligned}$$

according to part (i) of this lemma. The proof of the case when $j = s-2, \cdots, 0$ are similar, so is

omitted.

To prove Lemma 1.8(viii), note that

$$\begin{aligned}\left\|\frac{\partial L\left(z,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right\| &\leq \int_{\mathcal{X}}\left\|G'\left(z-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\mathbf{X}|z}\left(\mathbf{X}|\,z,\boldsymbol{\beta}\right)\mathbf{X}\right\|d\mathbf{X}\\ &\quad+\int_{\mathcal{X}}\left\|G\left(Z-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\frac{\partial f_{\mathbf{X}|z}\left(\mathbf{X}|\,z,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right\|d\mathbf{X}.\end{aligned}$$

Obviously, the first term on the RHS is bounded by $\left\|G'\right\|_{\infty}\sqrt{p}$, and the second term is bounded by $\left\|G\right\|_{\infty}\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X}|z}\left(\mathbf{X}|\,z,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\mathbf{X}$. Note that

$$\begin{aligned}\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X}|z}\left(\mathbf{X}|\,z,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\mathbf{X} &\leq \frac{2\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\mathbf{X}}{\int_{\mathcal{X}}f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)d\mathbf{X}}\\ &\leq \frac{2C\sqrt{p}m\left(\mathbb{X}\left(z,\boldsymbol{\beta}\right)\right)}{\zeta^{-1}m\left(\mathbb{X}\left(z,\boldsymbol{\beta}\right)\right)}\leq C\sqrt{p},\end{aligned}$$

according to part (iii) of this lemma. This proves Lemma 1.8(viii). Lemma 1.8(ix) can be similarly proved.

Finally, to show Lemma 1.8(x), we note that

$$\begin{aligned}&\int_{\mathcal{X}}\left\|\frac{\partial W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right\|d\widetilde{\boldsymbol{X}}\\ &\leq\int_{\mathcal{X}}\left\|G''\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^{\star}\right)+\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)\right\|f_{\mathbf{X}|z}\left(\widetilde{\mathbf{X}}\middle|\,z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)d\widetilde{\mathbf{X}}\\ &+\int_{\mathcal{X}}\left|G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^{\star}\right)+\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\right|\left\|\frac{\partial f_{\mathbf{X}|z}\left(\widetilde{\mathbf{X}}\middle|\,z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right\|d\widetilde{\mathbf{X}}.\end{aligned}$$

Obviously, the first term is bounded by $2\sqrt{p}\left\|G''\right\|_{\infty}$, and the second term is bounded by $\left\|G'\right\|_{\infty}\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X}|Z}\left(\widetilde{\mathbf{X}},z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|\boldsymbol{\beta}\right.$ Note that

$$\int_{\mathcal{X}}\left\|\frac{\partial f_{\mathbf{X}|z}\left(\widetilde{\mathbf{X}}\middle|\,z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right\|d\widetilde{\mathbf{X}}\leq\frac{2\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}},z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\widetilde{\mathbf{X}}}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}.$$

We can see that

$$\frac{\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} = \left.\frac{\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\middle|\boldsymbol{\beta}\right)}{\partial z}\right|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})} \mathbf{X} + \left.\frac{\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\middle|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}\right|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})},$$

according to (i) and (ii), we know that $\left\|\left.\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\middle|\boldsymbol{\beta}\right)/\partial z\right|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}\right\|$ is bounded, and $\left\|\left.\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\middle|\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}\right\|$ is bounded by $C\sqrt{p}$, so $\left\|\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|$ is bounded by $C\sqrt{p}$. So

$$\frac{\int_{\mathcal{X}}\left\|\partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\| d\widetilde{\mathbf{X}}}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\boldsymbol{\beta}\right)} \leq \frac{C\sqrt{p}\cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)}{\zeta^{-1}\cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)} = C\sqrt{p}.$$

This finishes the proof of Lemma 1.8(xii). □

Lemma 1.9. *Suppose that Assumption 2.1, Assumption 2.2(i)-(iii), 2.3 and Assumption 2.4 hold with $\upsilon_G = 3$, $\upsilon_K = 2$, and $\upsilon_f = 3$. Define*

$$A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) = \frac{1}{nh_n}\sum_{j=1}^{n} K\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\cdot\left(\cdot_j\right),$$

where $\cdot$ is y or 1. Also define $A_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) = \lim_{n\to\infty}\mathbb{E}_{\mathscr{D}_n}A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$, where the expectation $\mathbb{E}_{\mathscr{D}_n}$ is taken with respect to the data set $\mathscr{D}_n$. Then

(i) There holds

$$\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\mathcal{X}_e\times\in\mathcal{B}}\left|A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - \mathbb{E}_{\mathscr{D}_n}A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right| = O_p\left(h_n^{-1}\sqrt{p\log\left(nph_n^{-1}\right)/n}\right);$$

(ii) There holds

$$\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\mathcal{X}_e\times\in\mathcal{B}}\left|\mathbb{E}_{\mathscr{D}_n}A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - A_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right| = O_p\left(h_n^2\right);$$

(iii) Define $\psi\left(n,p,h_n\right) = h_n^{-1}\sqrt{p\log\left(nph_n^{-1}\right)/n} + h_n^2$, there holds

$$\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\mathcal{X}_e\times\in\mathcal{B}}\left|A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - A_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right| = O_p\left(h_n^{-1}\sqrt{p\log\left(nph_n^{-1}\right)/n} + h_n^2\right).$$

Proof. Lemma 3.5(i) is a direct result of Lemma 1.7 if we note that

$$\left|K\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\cdot\left(\cdot_j\right)\right|\leq Ch_h^{-1}$$

and

$$\left\|\partial\left(K\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\cdot\left(\cdot_j\right)\right)/\partial\boldsymbol{\beta}\right\|\leq C\sqrt{p}h_h^{-2}.$$

To prove Lemma 3.5(ii), we only need to note that

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}_n}\left[A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right]\\
&=\frac{1}{h_n}\mathbb{E}_{\mathscr{D}_n}\left[K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)}{h_n}\right)y_j\right]\\
&=\frac{1}{h_n}\mathbb{E}_{\mathscr{D}_n}\left[K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)}{h_n}\right)G\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)-\mathbf{X}_j^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\right]\\
&=\frac{1}{h_n}\int K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z}{h_n}\right)G\left(z-\mathbf{X}_j^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\mathbf{X},z}\left(\mathbf{X}_j,z|\,\boldsymbol{\beta}\right)d\mathbf{X}_j dz\\
&=\frac{1}{h_n}\int K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z}{h_n}\right)f_z\left(z|\,\boldsymbol{\beta}\right)dz\int_{\mathcal{X}}G\left(z-\mathbf{X}_j^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\frac{f_{\mathbf{X},z}\left(\mathbf{X}_j,z|\,\boldsymbol{\beta}\right)}{f_z\left(z|\,\boldsymbol{\beta}\right)}d\mathbf{X}_j\\
&=\frac{1}{h_n}\int K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z}{h_n}\right)f_z\left(z|\,\boldsymbol{\beta}\right)L\left(z,\boldsymbol{\beta}\right)dz\\
&=\int K\left(z\right)L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-h_nz,\boldsymbol{\beta}\right)f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-h_nz|\,\boldsymbol{\beta}\right)dz\\
&=L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)+\frac{h_n^2}{2}\left[\frac{\partial^2L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial z^2}\right]\left[\int K\left(z\right)z^2dz\right]\\
&+\frac{h_n^3}{6}\left\{\int K\left(z\right)z^3\left[\frac{\partial^3L\left(\widetilde{z},\boldsymbol{\beta}\right)f_z\left(\widetilde{z}|\,\boldsymbol{\beta}\right)}{\partial z^3}\right]dz\right\},
\end{aligned}$$

and similarly,

$$\begin{aligned}
\mathbb{E}_{\mathscr{D}_n}\left[A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right]&=\frac{1}{h_n}\mathbb{E}_{\mathscr{D}_n}\left[K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)}{h_n}\right)\right]\\
&=\frac{1}{h_n}\int\left[K\left(\frac{z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z}{h_n}\right)f_z\left(z|\,\boldsymbol{\beta}\right)\right]dz\\
&=\int K\left(z\right)f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-h_nz|\,\boldsymbol{\beta}\right)dz\\
&=f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)+\frac{h_n^2}{2}\left[\frac{\partial^2f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial z^2}\right]\left[\int K\left(z\right)z^2dz\right]\\
&+\frac{h_n^3}{6}\left\{\int K\left(z\right)z^3\left[\frac{\partial^3f_z\left(\widetilde{z}|\,\boldsymbol{\beta}\right)}{\partial z^3}\right]dz\right\},
\end{aligned}$$

where $\widetilde{z}$ lies between $z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$ and z. Note that according to Lemma 1.8 (i) and (ii), $f_z\left(z|\,\boldsymbol{\beta}\right)$ and $L\left(z,\boldsymbol{\beta}\right)f_z\left(z|\,\boldsymbol{\beta}\right)=\int_{\mathcal{X}}G\left(z-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)d\mathbf{X}$ both have up to third bounded derivatives

with respect to z, so the results hold.

Finally, Lemma 3.5 (iii) is a combination of Lemma 3.5 (i) and Lemma 3.5 (ii). □

Lemma 1.10. *Suppose that Assumption 2.1, Assumption 2.2(i)-(iii), Assumption 2.3, and Assumption 2.4 hold. Given any positive sequence* $\{\phi_n\}_{n=1}^{\infty}$ *satisfying* $p\phi_n \downarrow 0$*, define*

$$\mathcal{X}_{e,n} = \{\boldsymbol{X}_e \in \mathcal{X}_e : |X_j| \leq 1 - \phi_n, 0 \leq j \leq p\}.$$

Then

(i) $1 - P(\mathbf{X}_e \in \mathcal{X}_{e,n}) = O(p\phi_n)$*, and* $\inf_{(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_{e,n} \times \mathcal{B}} f_Z(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta}) \sim \phi_n^p p^{-p}$*;*

(ii) If $\psi(n,p,h_n) = o(\phi_n^p p^{-p})$*, there holds*

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_{e,n} \times \mathcal{B}} \left| \widehat{G}(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta}) - L(z(\mathbf{X}_e,\boldsymbol{\beta}),\boldsymbol{\beta}) \right| = O_p\left(p^p \phi_n^{-p} \psi(n,p,h_n)\right).$$

Proof. To prove Lemma 3.6(i), note that for $p\phi_n < 1$, $m(\mathcal{X}_e - \mathcal{X}_{e,n}) = 1 - (1-\phi_n)^p \leq p\phi_n$. So $\int_{\mathcal{X}_e - \mathcal{X}_{e,n}} f_e(\mathbf{X}_e)\, d\mathbf{X}_e \leq \zeta p\phi_n = O(p\phi_n)$ due to Assumption 2.4(i). To show the lower bound, note that given any $\boldsymbol{\beta} \in \mathcal{B}$ and $\mathbf{X}_e \in \mathcal{X}_{e,n}$, there holds $|z(\mathbf{X}_e,\boldsymbol{\beta}) - \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta} - X_0| \leq \sum_{j=1}^p |\beta_j|\,|X_j - \widetilde{X}_j|$. This implies that for any $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{X}} \in \mathbb{X}(z(\mathbf{X}_e,\boldsymbol{\beta}),\boldsymbol{\beta})$ if

$$\widetilde{\mathbf{X}} \in \left\{ \widetilde{\mathbf{X}} \in [0,1]^p : \left(\sup_{\boldsymbol{\beta} \in \mathcal{B}} |\beta_j| \right) \left| X_j - \widetilde{X}_j \right| \leq \phi_n / p \right\}.$$

Since the above set has Lebesgue measure of order $O(\phi_n^p / p^p)$, we have that

$$\begin{aligned} &\inf_{(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_{e,n} \times \mathcal{B}} f_z(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta}) \\ &\geq \inf_{(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_{e,n} \times \mathcal{B}} \int_{\widetilde{\mathbf{X}} \in \mathbb{X}(z(\mathbf{X}_e,\boldsymbol{\beta}),\boldsymbol{\beta})} f_e\left(z(\mathbf{X}_e,\boldsymbol{\beta}) - \widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}, \widetilde{\mathbf{X}}\right) d\widetilde{\mathbf{X}} \sim \phi_n^p / p^p, \end{aligned}$$

due to Assumption 2.4(i). This proves Lemma 3.6(i).

To prove Lemma 3.6(ii), note that for any $\mathbf{X}_e$ and $\boldsymbol{\beta}$, we have $\widehat{G}(z(\mathbf{X}_e,\boldsymbol{\beta})|\boldsymbol{\beta}) = A_{n,y}(\mathbf{X}_e,\boldsymbol{\beta}) / A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta})$

and $L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)=A_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)/A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$. So

$$\begin{aligned}
&\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}\left|\widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)-L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right| \\
&\leq \sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}\frac{\left|A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} \\
&+\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\frac{\left|A_{n,1}\left(\mathbf{X},\boldsymbol{\beta}\right)-A_1\left(\mathbf{X},\boldsymbol{\beta}\right)\right|}{A_1\left(\mathbf{X},\boldsymbol{\beta}\right)}.
\end{aligned}$$

Obviously, since $\psi_1\left(n,p,h_n\right)=o\left(\phi_n^p/p^p\right)$,

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}\left|A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|=o_p\left(\phi_n^p/p^p\right),$$

so $\inf_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}A_{n,1}^{-1}\left(\boldsymbol{X}_e,\boldsymbol{\beta}\right)=O_p\left(p^p\phi_n^{-p}\right)$. Moreover, $L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)$ is upper bounded by Lemma 1.8(vii). Then the results hold according to Lemma 3.5. □

Proof of Lemma 1.1.

Proof. Note that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\| \\
&\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)\mathbf{X}_i\right\| \qquad (1.1)\\
&+\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|. \qquad (1.2)
\end{aligned}$$

Obviously, (1.1) is bounded by

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)\mathbf{X}_i\right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot I_{n,i} \qquad (1.3)\\
&+\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot\left(1-I_{n,i}\right), \qquad (1.4)
\end{aligned}$$

where $I_{n,i} = I\left(\mathbf{X}_{e,i} \in \mathcal{X}_{e,n}\right)$ and $\mathcal{X}_{e,n}$ is chosen as in Lemma 3.6. Note that (1.3) is bounded by

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i - L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot I_{n,i}\\
&\leq \sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X} - L\left(Z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}\right\|\\
&= O_p\left(p^{p+1/2}\phi_n^{-p}\psi_1\left(n,p,h_n\right)\right),
\end{aligned}$$

according to Lemma 3.6. For (1.4), we have that

$$\begin{aligned}
&\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i - L\left(Z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot\left(1-I_{n,i}\right)\\
&\leq C\sqrt{p}\mathbb{E}I\left(\mathbf{X}_{e,i}\notin\mathcal{X}_{e,n}\right) = O\left(p^{3/2}\phi_n\right),
\end{aligned}$$

according to Lemma 3.6(i). Then we have that (1.3) is of order $O_p\left(p^{p+1/2}\phi_n^{-p}\psi_1\left(n,p,h_n\right)+p^{3/2}\phi_n\right)$.

Now we go to (1.2). Similar to the above truncation, we have that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|\\
&\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot I_{n,i} - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot I_{n,i}\right]\right\| \qquad (1.5)\\
&+\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot\left(1-I_{n,i}\right) - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot\left(1-I_{n,i}\right)\right]\right\|. \qquad (1.6)
\end{aligned}$$

Obviously, (1.6) is $O_p\left(p^{3/2}\phi_n\right)$. For (1.5), note that $\left\|L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)X_{j,i}\cdot I_{n,i}\right\|$ is bounded by C and $\partial\left\|L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)X_{j,i}\cdot I_{n,i}/\partial\boldsymbol{\beta}\right\|$ is bounded by $C\sqrt{p}$ by Lemma 1.8(vii) and (viii), we have that (1.5) is of order $O_p\left(\sqrt{p^2 n\log(pn)/n}\right)$ using Lemma 1.7. Then

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|\\
&= O_p\left(\sqrt{p^2\log\left(pn\right)/n}+p^{3/2}\phi_n\right).
\end{aligned}$$

Together, we have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|$$
$$=O_p\left(p^{p+1/2}\phi_n^{-p}\psi_1\left(n,p,h_n\right)+\sqrt{p^2\log\left(pn\right)/n}+p^{3/2}\phi_n\right).$$

Then if we set $\phi_n=p^{\frac{p-1}{p+1}}\psi_1^{\frac{1}{p+1}}\left(n,p,h_n\right)$, we have that

$$p\phi_n=p^p\phi_n^{-p}\psi_1\left(n,p,h_n\right)=p^{\frac{2p}{p+1}}\psi_1^{\frac{1}{p+1}}\left(n,p,h_n\right)\leq p^{\frac{5p+1}{2(p+1)}}\psi_1^{\frac{1}{p+1}}\left(n,p,h_n\right)\to 0,$$

and

$$\sqrt{p^2\log\left(pn\right)/n}=o\left(p^{\frac{5p+1}{2(p+1)}}\psi_1^{\frac{1}{p+1}}\left(n,p,h_n\right)\right),$$

so

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|=O_p\left(p^{\frac{5p+1}{2(p+1)}}\psi_1^{\frac{1}{p+1}}\left(n,p,h_n\right)\right).$$

This finishes the whole proof. □

Lemma 1.11. *Suppose that p is fixed. If all the assumptions in Lemma 3.5 hold with $\upsilon_G=4$, $\upsilon_K=3$, and $\upsilon_f=4$, we have that Lemma 3.5(i) holds. Moreover,*

(i) There holds

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e\times\in\mathcal{B}}\left|\mathbb{E}_{\mathscr{D}_n}A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|=O_p\left(h_n^3\right);$$

(ii) There holds

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e\times\in\mathcal{B}}\left|A_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right|=O_p\left(h^{-1}\sqrt{\log\left(nh^{-1}\right)/n}+h^3\right).$$

Proof. The proof is similar to the proof of Lemma 3.5 so is omitted. □

Lemma 1.12. *Suppose that p is fixed. For any $\mathbf{X}_e\in\mathcal{X}_e$ and $\boldsymbol{\beta}\in\mathcal{B}$, define*

$$A'_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)=\frac{1}{nh_n^2}\sum_{j=1}^{n}K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\left(\mathbf{X}-\mathbf{X}_j\right)\cdot\left(\cdot_j\right),$$

where $\cdot=1$ or $\cdot=y$. If all the assumptions in Lemma 3.5 hold with $\upsilon_G=4$, $\upsilon_K=3$, and $\upsilon_f=4$, then

(i) There holds

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e\times\mathcal{B}}\left\|A'_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\mathbb{E}_{\mathscr{D}_n}A'_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right\|=O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}\right);$$

(ii) Define $A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)=\lim_{n\to\infty}\mathbb{E}_{\mathscr{D}_n}A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$ *and* $A'_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)=\lim_{n\to\infty}\mathbb{E}_{\mathscr{D}_n}A'_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$. *We have that* $A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)=\partial H_1\left(z,\mathbf{X}|\,\boldsymbol{\beta}\right)/\partial z|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}$ *and* $A'_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)=\partial H_2\left(z,\mathbf{X}|\,\boldsymbol{\beta}\right)/\partial z|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}$, *where*

$$H_1\left(z,\mathbf{X}|\,\boldsymbol{\beta}\right)=\int_{\mathcal{X}}G\left(z-\widetilde{\mathbf{X}}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_e\left(z-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}},$$

$$H_2\left(z,\mathbf{X}|\,\boldsymbol{\beta}\right)=\int_{\mathcal{X}}f_e\left(z-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}},$$

and the differentiation of H_1 *and* H_2 *are element-wise. Moreover, there holds*

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e\times\mathcal{B}}\left\|\mathbb{E}_{\mathscr{D}_n}A'_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A'_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right\|=O_p\left(h_n^3\right),$$

(iii) There holds

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e\times\mathcal{B}}\left\|A'_{n,\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-A'_{\cdot}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right\|=O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3\right).$$

Proof. Lemma 3.8(i) is a direct result of Lemma 1.7 if we note that for each $1\leq l\leq p$, $h_n^{-2}K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\left(\cdot_j\right)$ is bounded by Ch_n^{-2} and its derivatives with respect to $\boldsymbol{\beta}$ and $\mathbf{X}$ are both upper bounded since p is fixed.

To prove Lemma 3.8(ii), we note that

$$\begin{aligned}
&\mathbb{E}_{\mathscr{D}_n}A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\\
&=\frac{1}{h_n^2}\mathbb{E}_{\mathscr{D}_n}\left[K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\left(\mathbf{X}-\mathbf{X}_j\right)\cdot G\left(X_{0,j}+\mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\right]\\
&=\frac{1}{h_n^2}\mathbb{E}_{\mathscr{D}_n}\left[K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\left(\mathbf{X}-\mathbf{X}_j\right)\cdot G\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)-\mathbf{X}_j^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\right]\\
&=\frac{1}{h_n^2}\int K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\right)/h_n\right)dz\int_{\mathcal{X}}\left[G\left(z-\widetilde{\mathbf{X}}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}},z\middle|\,\boldsymbol{\beta}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)\right]d\widetilde{\mathbf{X}}\\
&=\frac{1}{h_n^2}\int\left[K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-z\right)/h_n\right)H_1\left(z,\mathbf{X}|\,\boldsymbol{\beta}\right)\right]dz\\
&=\frac{1}{h_n}\int\left[K'\left(z\right)H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-h_nz,\mathbf{X}|\,\boldsymbol{\beta}\right)\right]dz
\end{aligned}$$

Note that both G and f_e have up to fourth bounded derivatives with respect to z, and the upper

bounds hold uniformly with respect to z, $\mathbf{X}$ and $\boldsymbol{\beta}$. This implies that each element of $H_1(z, \mathbf{X}|\boldsymbol{\beta})$ has up to fourth bounded derivatives with respect to z. Also ote that $\int K'(v)\,dv = K(v)|_{-\infty}^{\infty} = 0$, $\int vK'(v)\,dv = K(v)|_{-\infty}^{\infty} - \int K(v)\,dv = -1$, $\int v^s K'(v)\,dv = v^s K(v)|_{-\infty}^{\infty} - s\int v^{s-1}K(v)\,dv = 0$ for $s = 2, 3$, and $\left|\int v^4 K'(v)\,dv\right| < \infty$. This implies that

$$\left\|\mathbb{E}_{\mathscr{D}_n} A'_{n,y}(\mathbf{X}_e, \boldsymbol{\beta}) - A'_y(\mathbf{X}_e, \boldsymbol{\beta})\right\| = O_p\left(h_n^3\right)$$

uniform with respect to $\mathbf{X}_e$ and $\boldsymbol{\beta}$. The proof of the uniform distance between $\mathbb{E}_{\mathscr{D}_n} A'_{n,1}(\mathbf{X}_e, \boldsymbol{\beta})$ and $A'_1(\mathbf{X}_e, \boldsymbol{\beta})$ is similar. So we finish the proof of Lemma 3.8(ii).

Finally, Lemma 3.8(iii) is a combination of Lemma 3.8(i) and Lemma 3.8(ii). □

Lemma 1.13. *Suppose that p is fixed. If all the assumptions in Lemma 3.5 hold with $\upsilon_G = 4$, $\upsilon_K = 3$, and $\upsilon_f = 4$, we have that*

$$\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_e^{\phi}\times\mathcal{B}} \left\| \frac{\partial \widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} - \frac{\partial H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)} + L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\frac{\partial H_2\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)} \right\| = O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right),$$

where $\mathcal{X}_e^{\phi}$ is defined in (1.13) in the main text.

Proof. Note that

$$\begin{aligned}\frac{\partial \widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} &= \frac{\partial A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} - \frac{A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\cdot\frac{\partial A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} \\ &= \frac{A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} - \frac{A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\frac{A'_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}.\end{aligned}$$

Then

$$\begin{aligned}\left\|\frac{A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} - \frac{\partial H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)}\right\| &= \left\|\frac{A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)} - \frac{A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\right\| \\ &\le \left\|\frac{A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\right\| \qquad (1.7) \\ &\quad + \left\|\frac{A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\frac{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\right\|. \qquad (1.8)\end{aligned}$$

Now for any $(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_e^{\phi}\times\mathcal{B}$, $A_1\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$ is uniformly lower-bounded according to Lemma 3.6, so $A_{n,1}^{-1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right) = O_p(1)$ also uniformly holds. Moreover, $\left\|A'_y\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right\|$ is upper bounded, so

$\left\|A'_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right\| = O_p\left(1\right)$ also uniformly holds. Then (1.7) is $O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3\right)$ and (1.8) is $O_p\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3\right)$. Similar method can be used to show that

$$\frac{A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}\frac{A'_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}{A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)}-\frac{L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\partial H_2\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)}$$

is also $O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3\right)$. This finishes the proof. □

Lemma 1.14. *Suppose that p is fixed. If all the assumptions in Lemma 3.5 hold with $\upsilon_G = 4$, $\upsilon_K = 3$, and $\upsilon_f = 4$, then for any $\overline{\mathcal{B}} \subseteq \mathcal{B}$, we have that*

$$\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\mathcal{X}_e^{\phi}\times\overline{\mathcal{B}}}\left\|\frac{\partial\widehat{G}\left(Z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}-\int W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}_e\right)d\widetilde{\mathbf{X}}_e\right\|\leq\alpha_{1,n}+\alpha_2,$$

where $\alpha_{1,n} = O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3\right)$ and $\alpha_2 = O_p\left(\sup_{\boldsymbol{\beta}\in\overline{\mathcal{B}}}\|\Delta\boldsymbol{\beta}\|\right)$.

Proof. We only need to show that

$$\begin{aligned}\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\overline{\mathcal{X}}_e\times\mathcal{B}}\Bigg\|&\frac{\partial H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)}-L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\frac{\partial H_2\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z}{f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)}\\&-\int W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\Bigg\|=O\left(\|\Delta\boldsymbol{\beta}\|\right).\end{aligned}$$

Note that

$$\begin{aligned}&\partial H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z-L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\partial H_2\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z\\&=\int G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}\Delta\boldsymbol{\beta}\right)f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\\&+\int G\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\partial f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)/\partial z\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\\&-L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\int\left(\partial f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)/\partial z\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\\&=\int G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\\&+\int\left[G\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)-G\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\right]\left(\partial f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)/\partial z\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}\\&-\left(L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)-G\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\right)\int\left(\partial f_e\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta},\widetilde{\mathbf{X}}\right)/\partial z\right)\left(\mathbf{X}-\widetilde{\mathbf{X}}\right)d\widetilde{\mathbf{X}}.\end{aligned}$$

Note that

$$\begin{aligned}
&\left\| \int \left[G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \Delta \boldsymbol{\beta} \right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right] \left(\partial f_e \left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \boldsymbol{\beta}, \widetilde{\mathbf{X}} \right) / \partial z \right) \left(\mathbf{X} - \widetilde{\mathbf{X}} \right) d\widetilde{\mathbf{X}} \right\| \\
&\leq C \cdot \sup_{\widetilde{\mathbf{x}} \in \mathcal{X}} \left| G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \Delta \boldsymbol{\beta} \right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right) \\
&\leq C \cdot \| \Delta \boldsymbol{\beta} \| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right),
\end{aligned}$$

and according to our choice of $\mathcal{X}_e^{\phi}$, we know that $m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right) > 0$. On the other side,

$$\begin{aligned}
&\left\| \left(L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right) \int \left(\partial f_e \left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \boldsymbol{\beta}, \widetilde{\mathbf{X}} \right) / \partial z \right) \left(\mathbf{X} - \widetilde{\mathbf{X}} \right) d\widetilde{\mathbf{X}} \right\| \\
&\leq C \cdot \left| L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right) \\
&= C \cdot \left| L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) - L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta}^{\star} \right) \right| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right) \\
&\leq C \cdot \left(\sup_{z, \boldsymbol{\beta}} \left\| \partial L\left(z, \boldsymbol{\beta} \right) / \partial \boldsymbol{\beta} \right\| \right) \cdot \| \Delta \boldsymbol{\beta} \| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right) \\
&\leq C \cdot \| \Delta \boldsymbol{\beta} \| \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right)
\end{aligned}$$

due to the upper boundedness of $\left\| \partial L\left(z, \boldsymbol{\beta} \right) / \partial \boldsymbol{\beta} \right\|$ according to Lemma 1.8(viii). Note that

$$f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) > C \cdot m\left(\mathbb{X}\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X} \right) \right)$$

for some $C > 0$ due to Assumption 2.4(i) and the choice of $\mathcal{X}_e^{\phi}$, so we have that

$$\begin{aligned}
&\left\| \left(\partial H_1\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X}_e \right) / \partial z - L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) \partial H_2\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X}_e \right) / \partial z \right) / f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) \right. \\
&\left. - \int G'\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \Delta \boldsymbol{\beta} \right) f_e \left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) - \widetilde{\mathbf{X}}^{\mathrm{T}} \boldsymbol{\beta}, \widetilde{\mathbf{X}} \right) \left(\mathbf{X} - \widetilde{\mathbf{X}} \right) d\widetilde{\mathbf{X}} / f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) \right\| \\
&= \left\| \left(\partial_z H_1\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X}_e \right) - L\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) \partial_z H_2\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right), \mathbf{X}_e \right) \right) / f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) \right. \\
&\left. - \int W\left(\mathbf{X}_e, \widetilde{\mathbf{X}}_e, \boldsymbol{\beta} \right) \left(\mathbf{X} - \widetilde{\mathbf{X}} \right) d\widetilde{\mathbf{X}} \right\| \leq C \cdot \| \Delta \boldsymbol{\beta} \|.
\end{aligned}$$

This proves the results. □

Now we prove Lemma 1.2 in the main text.

Proof of Lemma 1.2. Note that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}-\Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right\| \\
\leq &\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\left(\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}-\int W\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)\left(\mathbf{X}_i-\mathbf{X}\right)d\mathbf{X}\right)\right\| \qquad (1.9)\\
+&\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\left(\int W\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)\left(\mathbf{X}_i-\mathbf{X}\right)d\mathbf{X}\right)-\Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right\|. \qquad (1.10)
\end{aligned}$$

Obviously, (1.9) is of order $O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n}+h_n^3+\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\|\Delta\boldsymbol{\beta}\|\right)$ according to Lemma 3.2. Using Lemma 1.7, we can show that (1.10) is $O_p\left(\sqrt{\left(\log n\right)/n}\right)$ by noting that each element of $\int W\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)\left(\mathbf{X}_i-\mathbf{X}\right)d\boldsymbol{X}$ is bounded and that $\int_{\mathcal{X}}\left\|\partial W\left(\mathbf{X}_e,\widetilde{\mathbf{X}}_e,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\widetilde{\mathbf{X}}$ is uniformly upper bounded according to Lemma 1.8(x). This finishes the proof of Lemma 1.2. □

Now we prove Lemma 1.3 in the main text.

Proof of Lemma 1.3. We first show that

$$\boldsymbol{\xi}_n^{\phi}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\left(\frac{y_j-y_i}{f_z^{\star}\left(z_i^{\star}\right)}\right)\mathbf{X}_i^{\phi}+o_p\left(\frac{1}{\sqrt{n}}\right),$$

Define $f_z^{\star}\left(z_i^{\star}\right)=f_z\left(z|\,\boldsymbol{\beta}^{\star}\right)$ and $f_{\mathbf{X},z}^{\star}\left(\mathbf{X},z\right)=f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}^{\star}\right)$. Recall that $z_i^{\star}=z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}^{\star}\right)$, so

$$\begin{aligned}
&\boldsymbol{\xi}_n^{\phi}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\left(\frac{y_j-y_i}{f_z^{\star}\left(z_i^{\star}\right)}\right)\mathbf{X}_i^{\phi}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\left(y_j-y_i\right)\right]\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)}-\frac{1}{f_z^{\star}\left(z_i^{\star}\right)}\right]\mathbf{X}_i^{\phi}\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\left(y_j-G\left(z_i^{\star}\right)\right)\right]\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)}-\frac{1}{f_z^{\star}\left(z_i^{\star}\right)}\right]\mathbf{X}_i^{\phi}(i)\\
&-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\right]\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)}-\frac{1}{f_z^{\star}\left(z_i^{\star}\right)}\right]\mathbf{X}_i^{\phi}(ii).
\end{aligned}$$

For term (i), due to truncation, we have that

$$\max_{1\leq i\leq n}\left\|\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)}-\frac{1}{f_z\left(z_i^{\star}\right)}\right]\mathbf{X}_i^{\phi}\right\|=O_p\left(h_n^{-1}\sqrt{\log\left(n\right)/n}+h_n^3\right).$$

We further provide a uniform bound for $\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^{\star}-z_i^{\star}\right)\left(y_j-G\left(z_i^{\star}\right)\right)\mathbf{X}_i^{\phi}$ over i. We first note

that

$$\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(y_j - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right] = \mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(G\left(z_j^\star\right) - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right],$$

where the RHS is equivalent to

$$\begin{aligned}
&\mathbb{E}\left\{\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(G\left(z_j^\star\right) - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\middle|\mathbf{X}_{e,i}\right]\right\} \\
&= \frac{n-1}{n}\mathbb{E}\left\{\mathbf{X}_i^\phi\int\left[K_{h_n}\left(z - z_i^\star\right)\left(G\left(z\right) - G\left(z_i^\star\right)\right)f_z^\star\left(z\right)\right]dz\right\} \\
&= \frac{n-1}{n}\mathbb{E}\left\{\mathbf{X}_i^\phi 0\int\left[K\left(z\right)\left(G\left(z_i^\star + zh_n\right) - G\left(z_i^\star\right)\right)f_z^\star\left(z_i + zh_n\right)\right]dz\right\}.
\end{aligned}$$

Now note that since G and $f_z^\star$ both have up to fourth order bounded derivatives, we have that

$$\begin{aligned}
&\left(G\left(z_i^\star + zh_n\right) - G\left(z_i^\star\right)\right)f_z^\star\left(z_i + zh_n\right) \\
&= \left(G'\left(z_i^\star\right)zh_n + \frac{1}{2}G''\left(z_i^\star\right)z^2h_n^2 + \frac{1}{6}G'''\left(z_i^\star\right)z^3h_n^3 + O\left(z^4h_n^4\right)\right)\left(f_z^\star\left(z_i^\star\right) + O\left(zh_n\right)\right) \\
&= G'\left(z_i^\star\right)f_z^\star\left(z_i^\star\right)zh_n + \frac{1}{2}G''\left(z_i^\star\right)f_z^\star\left(z_i^\star\right)z^2h_n^2 + \frac{1}{6}G'''\left(z_i^\star\right)f_z^\star\left(z_i^\star\right)z^3h_n^3 + O\left(z^4h_n^4\right).
\end{aligned}$$

So

$$\int\left[K\left(z\right)\left(G\left(z_i^\star + zh_n\right) - G\left(z_i^\star\right)\right)f_z^\star\left(z_i + zh_n\right)\right]dz = O\left(h_n^3\right),$$

where the bound does not depend on i. So

$$\max_{1\leq i\leq n}\left\|\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(G\left(z_j^\star\right) - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right]\right\| = O\left(h_n^3\right).$$

On the other side, we have that we have that

$$\begin{aligned}
\max_{1\leq i\leq n}\Bigg\|&\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(y_j - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi \\
&-\mathbb{E}_{\mathscr{D}_n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)\left(y_j - G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right]\Bigg\| = O_p\left(\sqrt{\left(\log n\right)/nh_n^2}\right).
\end{aligned}$$

Together we have that

$$\max_{1\leq i\leq n}\left\|\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)\left(y_j-G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right\|=O_p\left(h_n^{-1}\sqrt{\left(\log n\right)/n}+h_n^3\right).$$

So

$$\begin{aligned}&\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)\left(y_j-G\left(z_i^\star\right)\right)\right]\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}-\frac{1}{f_z\left(z_i^\star\right)}\right]\mathbf{X}_i^\phi\right\|\\&\leq\max_{1\leq i\leq n}\left\|\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)\left(y_j-G\left(z_i^\star\right)\right)\mathbf{X}_i^\phi\right\|\max_{1\leq i\leq n}\left|\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}-\frac{1}{f_z\left(z_i^\star\right)}\right|\\&=O_p\left(h_n^{-2}\left(\log n\right)/n+h_n^6\right)=o_p\left(1/\sqrt{n}\right),\end{aligned}$$

according to our choice of h_n, so term (i) is $o_p\left(1/\sqrt{n}\right)$.

For term (ii), without of loss of generality, we assume that $\mathbf{X}_i^\phi=X_i^\phi$ is a scalar; the general case can be proved similarly. We note that

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{n}\varepsilon_i\left[\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)\right]\left[\frac{1}{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}-\frac{1}{f_z^\star\left(z_i^\star\right)}\right]X_i^\phi\right]\\&=\mathbb{E}\sum_{i=1}^{n}\mathbb{E}\left\{\varepsilon_i\left[1-\frac{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}{f_z^\star\left(z_i^\star\right)}\right]X_i^\phi\middle|X_i\right\}=0\end{aligned}$$

due to the fact that the data is i.i.d. and that $\mathbb{E}\left(\varepsilon_i\middle|\mathbf{X}_{e,i}\right)=0$ for all i. Moreover,

$$\begin{aligned}&\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left[1-\frac{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}{f_z^\star\left(z_i^\star\right)}\right]X_i^{\phi2}\right]\\&=\frac{1}{n}\mathbb{E}\left\{G\left(z_i^\star\right)\left(1-G\left(z_i^\star\right)\right)\left[1-\frac{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)}{f_z^\star\left(z_i^\star\right)}\right]^2X_i^{\phi2}\right\}\\&\leq\frac{C}{n}\mathbb{E}\left\{\left(\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star-z_i^\star\right)-f_z^\star\left(z_i^\star\right)\right)^2X_i^{\phi2}\right\}\\&=\frac{C}{n^3}\mathbb{E}X_i^{\phi2}\left(\sum_{j\neq i,k\neq i,j\neq k}^{n}\mathbb{E}\left[\left(K_{h_n}\left(z_j^\star-Z_i^\star\right)-f_z^\star\left(z_i^\star\right)\right)\left(K_{h_n}\left(z_k^\star-z_i^\star\right)-f_z^\star\left(z_i^\star\right)\right)\middle|X_i^\phi\right]+O\left(nh_n^{-1}\right)\right)\end{aligned}$$

Note that $\mathbb{E}\left[\left(K_{h_n}\left(z_j^\star-Z_i^\star\right)-f_z^\star\left(z_i^\star\right)\right)\left(K_{h_n}\left(z_k^\star-z_i^\star\right)-f_z^\star\left(z_i^\star\right)\right)\middle|X_i^\phi\right]$ is $O\left(h_n^6\right)$ for all $k\neq j$, $j\neq i$,

and $k \neq i$. So the above term is of order $O\left(h_n^6/n + h_n^{-1}/n^2\right)$, implying that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left[1 - \frac{\frac{1}{n} \sum_{j=1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right)}{f_z^\star\left(z_i^\star\right)} \right] \mathbf{X}_i^\phi \right\| = O_p\left(h_n^3/\sqrt{n} + 1/\left(n\sqrt{h_n}\right)\right) = o_p\left(1/\sqrt{n}\right),$$

according to the choice of h_n. This proves the first result.

Now we obtain the asymptotic distribution of

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right) \left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \right) \mathbf{X}_i^\phi.$$

First note that

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right) \left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \right) \mathbf{X}_i^\phi \\
&= \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right) \left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)} \mathbf{X}_j^\phi \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right) \left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)} \mathbf{X}_j^\phi \right) \\
&= \frac{n\left(n-1\right)}{2n^2} \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right) \left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)} \mathbf{X}_j^\phi \right).
\end{aligned}$$

Let $\mathbb{E}_{j|i}$ be the expectation with respect to the j-th observation conditional on the i-th observation. Note that

$$\begin{aligned}
&\mathbb{E}_{j|i}\left[K_{h_n}\left(z_j^\star - z_i^\star\right) \frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)} \mathbf{X}_i^\phi \right] \\
&= \frac{\mathbf{X}_i^\phi}{f_z^\star\left(z_i^\star\right)} \mathbb{E}_{j|i}\left[K_{h_n}\left(z_j^\star - z_i^\star\right) \left(G\left(z_j^\star\right) - y_i \right) \right] \\
&= \frac{\mathbf{X}_i^\phi}{f_z^\star\left(z_i^\star\right)} \int K\left(z\right) \left(G\left(z_i^\star + h_n z\right) - y_i \right) f_z^\star\left(z_i^\star + h_n z\right) dz \\
&= \frac{\mathbf{X}_i^\phi}{f_z^\star\left(z_i^\star\right)} \int K\left(z\right) \left(G\left(z_i^\star\right) + G'\left(z_i^\star\right) z h_n + \frac{1}{2} G''\left(z_i^\star\right) z^2 h_n^2 + O\left(z^3 h_n^3\right) - y_i \right) \left(f_z^\star\left(z_i^\star\right) + O\left(z h_n\right) \right) dz \\
&= \frac{\mathbf{X}_i^\phi}{f_z^\star\left(z_i^\star\right)} \int K\left(z\right) \left(G\left(z_i^\star\right) - y_i \right) f_z^\star\left(z_i^\star\right) dz + O\left(h_n^3\right) = \mathbf{X}_i^\phi \left(G\left(z_i^\star\right) - y_i \right) + O\left(h_n^3\right),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{j|i}\left[K_{h_n}\left(z_j^\star - z_i^\star\right)\left(\frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)}\right)\mathbf{X}_j^\phi\right] \\
&= \int \frac{1}{h_n} K\left(\frac{z - z_i^\star}{h_n}\right)\left(\frac{y_i - G\left(z\right)}{f_z^\star\left(z\right)}\right)\mathbf{X}^\phi f_{\mathbf{X},z}^\star\left(\mathbf{X}, z\right) dz d\mathbf{X} \\
&= \int K\left(z\right)\frac{y_i - G\left(z_i^\star + h_n z\right)}{f_z^\star\left(z_i^\star + h_n z\right)}\mathbf{X}^\phi f_{\mathbf{X},z}^\star\left(\mathbf{X}, z_i^\star + h_n z\right) dz d\mathbf{X} \\
&= \left(y_i - G\left(z_i^\star\right)\right)\int \mathbf{X}^\phi f_{\mathbf{X}|z}^\star\left(\mathbf{X}\middle| z_i^\star\right) d\mathbf{X} + O\left(h_n^2\right) \\
&= \left(y_i - G\left(z_i^\star\right)\right)\mathbb{E}\left(\mathbf{X}^\phi\middle| z_i^\star\right) + O\left(h_n^3\right).
\end{aligned}$$

So

$$\mathbb{E}_{j|i}\left[K_{h_n}\left(z_j^\star - z_i^\star\right)\left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)}\mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)}\mathbf{X}_j^\phi\right)\right] = -\varepsilon_i\left(\mathbf{X}_i^\phi - \mathbb{E}\left(\mathbf{X}_i^\phi\middle| z_i^\star\right)\right) + O\left(h_n^3\right).$$

We also note that

$$\mathbb{E}\left\|K_{h_n}\left(z_j^\star - z_i^\star\right)\left(\frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)}\right)\mathbf{X}_j^\phi\right\|^2 \leq C\mathbb{E}\left(K_{h_n}^2\left(z_j^\star - z_i^\star\right)\right) = O\left(h_n^{-2}\right) = o\left(n\right),$$

$$\mathbb{E}_i\mathbb{E}_{j|i}\left[K_{h_n}\left(z_j^\star - z_i^\star\right)\left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)}\mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)}\mathbf{X}_j^\phi\right)\right] = O\left(h_n^3\right) = o\left(\frac{1}{\sqrt{n}}\right),$$

so according to Powell et al. (1989), we have that

$$\begin{aligned}
&\sqrt{n}\binom{n}{2}^{-1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K_{h_n}\left(z_j^\star - z_i^\star\right)\left(\frac{y_j - y_i}{f_z^\star\left(z_i^\star\right)}\mathbf{X}_i^\phi + \frac{y_i - y_j}{f_z^\star\left(z_j^\star\right)}\mathbf{X}_j^\phi\right). \\
&= -\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left(\mathbf{X}_i^\phi - \mathbb{E}\left(\mathbf{X}_i^\phi\middle| z_i^\star\right)\right) + o_p\left(1\right).
\end{aligned}$$

This implies that

$$\sqrt{n}\boldsymbol{\xi}_n^\phi = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left(\mathbf{X}_i^\phi - \mathbb{E}\left(\mathbf{X}_i^\phi\middle| z_i^\star\right)\right) + o_p\left(1\right) \to_d N\left(0, \Sigma_{\boldsymbol{\xi}}^\phi\right).$$

□

Lemma 1.15. *Suppose that Assumption 2.1, Assumption 2.2(i) and (ii), and Assumption 1.6 hold,*

we have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}\left(\boldsymbol{\beta}\right)-\Gamma_{q}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\chi_{1,n}\right).$$

Proof of Lemma 1.15. This is a direct result of Lemma 1.7 by noting that $\left|r_{s}\left(z\right)r_{s}\left(z\right)\right|\leq D_{q,0}^{2}$ and $\left\|\partial\left(r_{s}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)r_{s}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)/\partial\boldsymbol{\beta}\right\|\leq C\sqrt{p}D_{q,0}D_{q,1}$. □

Lemma 1.16. *Suppose that Assumption 2.1, Assumption 2.2(i) and (ii), and Assumption 1.6 hold, and $\chi_{1,n}\to 0$ as $n\to\infty$. We have that*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)-\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\chi_{1,n}\right).$$

Proof of Lemma 1.16. First note that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|\underline{\lambda}\left(\Gamma_{q,n}\left(\boldsymbol{\beta}\right)\right)-\underline{\lambda}\left(\Gamma_{q}\left(\boldsymbol{\beta}\right)\right)\right|\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}\left(\boldsymbol{\beta}\right)-\Gamma_{q}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\chi_{1,n}\right),$$

and

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|\overline{\lambda}\left(\Gamma_{q,n}\left(\boldsymbol{\beta}\right)\right)-\overline{\lambda}\left(\Gamma_{q}\left(\boldsymbol{\beta}\right)\right)\right|\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}\left(\boldsymbol{\beta}\right)-\Gamma_{q}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\chi_{1,n}\right).$$

Since $\chi_{1,n}\to 0$, we have that with probability going to 1, there holds

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_{q,n}\left(\boldsymbol{\beta}\right)\right)\leq\frac{3\overline{\lambda}_{\Gamma}}{2},\ \inf_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_{q,n}\left(\boldsymbol{\beta}\right)\right)\geq\frac{\underline{\lambda}_{\Gamma}}{2},$$

indicating that $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)\right)=O_{p}\left(1\right)$.

Note that for any positive semi-definite matrices A and B, there holds $\min\left\{\underline{\lambda}_{A}\left\|B\right\|,\underline{\lambda}_{B}\left\|A\right\|\right\}\leq\left\|AB\right\|\leq\max\left\{\overline{\lambda}_{A}\left\|B\right\|,\overline{\lambda}_{B}\left\|A\right\|\right\}$, so we have that

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)-\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right\|=\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)\left(\Gamma_{q,n}\left(\boldsymbol{\beta}\right)-\Gamma_{q}\left(\boldsymbol{\beta}\right)\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right\|\\&\leq\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)\right)\right)\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right)\right)\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Gamma_{q,n}\left(\boldsymbol{\beta}\right)-\Gamma_{q}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\chi_{1,n}\right).\end{aligned}$$

□

Lemma 1.17. *Suppose that Assumption 2.1, Assumption 2.2(i) and (ii), and Assumption 1.6 hold, and moreover $\chi_{1,n}\to 0$ as $n\to\infty$. Define*

$$\mathcal{Z}=\left\{z:z=X_{0}+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\text{ for some }\mathbf{X}_{e}\in\mathcal{X}_{e}\text{ and }\boldsymbol{\beta}\in\mathcal{B}\right\}.$$

We have that

$$\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)-\mathfrak{X}_{q}\left(z,\boldsymbol{\beta}\right)\right\|=O_{p}\left(\sqrt{p}qD_{q,0}^{2}\chi_{1,n}\right).$$

Proof of Lemma 1.17. Note that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)-\mathfrak{X}_{q}\left(z,\boldsymbol{\beta}\right)\right\| \\
&\leq\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(z\right)\mathbf{X}_{i}\right)\right\| \\
&+\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q}\left(z,\boldsymbol{\beta}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(z\right)\mathbf{X}_{i}\right)\right\|.
\end{aligned}$$

For the first term, we have that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(z\right)\mathbf{X}_{i}\right)\right\| \\
&=\frac{1}{n}\sum_{i=1}^{n}\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\left(\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)-\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right)\boldsymbol{r}_{q}\left(z\right)\mathbf{X}_{i}\right\| \\
&\leq C\sqrt{p}qD_{q,0}^{2}\left\|\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)-\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\sqrt{p}qD_{q,0}^{2}\chi_{1,n}\right).
\end{aligned}$$

For the second term, we note that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q}\left(z,\boldsymbol{\beta}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(z\right)\mathbf{X}_{i}\right)\right\| \\
&\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}}\sup_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}}\sup_{\mathbf{X}_{e}\in\mathcal{X}_{e}}\left\|\mathfrak{X}_{q}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}},\boldsymbol{\beta}\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)\mathbf{X}_{i}\right)\right\|,
\end{aligned}$$

where uniformly for all $\boldsymbol{\beta},\boldsymbol{\beta}_{1},\boldsymbol{\beta}_{2},\widetilde{\boldsymbol{\beta}}\in\mathcal{B}$, $\mathbf{X}_{e}\in\mathcal{X}_{e}$, and $\mathbf{X}_{i}\in\mathcal{X}$, there hold

$$\left|\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)X_{i,j}\right|\leq CqD_{q,0}^{2},$$

and

$$\left\|\frac{\partial\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)X_{i,j}}{\partial\mathbf{X}_{e}}\right\|\leq C\sqrt{p}qD_{q,0}D_{q,1},$$

$$\left\|\frac{\partial\boldsymbol{r}_{q}^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_{q}\left(X_{0}+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)X_{i,j}}{\partial\widetilde{\boldsymbol{\beta}}}\right\|\leq C\sqrt{p}qD_{q,0}D_{q,1},$$

$$
\begin{aligned}
&\left\|\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)\boldsymbol{r}_q\left(X_0+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)-\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_2\right)\boldsymbol{r}_q\left(X_0+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)\right\| \\
&\leq\left\|\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)-\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)\boldsymbol{r}_q\left(X_0+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)\right\| \\
&+\left\|\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\left(\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)-\Gamma_q^{-1}\left(\boldsymbol{\beta}_2\right)\right)\boldsymbol{r}_q\left(X_0+\mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)\right\| \\
&\leq C\sqrt{p}qD_{q,0}D_{q,1}\left\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right\|+CqD_{q,0}^2\left\|\Gamma_q\left(\boldsymbol{\beta}_1\right)-\Gamma_q\left(\boldsymbol{\beta}_2\right)\right\|\leq C\sqrt{p}q^2D_{q,0}^3D_{q,1}\left\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right\|.
\end{aligned}
$$

So we have that the second term is of order $O_p\left(\sqrt{p}\chi_{1,n}\right)$. This finishes the proof. □

Lemma 1.18. *Suppose that Assumption 2.1, Assumption 2.2(i)-(iii), and Assumption 1.6 hold with $\upsilon_G\geq 1$, and that $\chi_{1,n}\to 0$ as $n\to\infty$, then we have that*

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^n\left(\mathbf{X}_i-\mathfrak{X}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}\right)\right)\left(G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)-G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star\right)\right)\right. \\
&\qquad\left.-\mathbb{E}\left(\left(\mathbf{X}_i-\mathfrak{X}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}\right)\right)\left(G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)-G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star\right)\right)\right)\right\|=O_p\left(\sqrt{p}\chi_{1,n}\right).
\end{aligned}
$$

Proof of Lemma 1.18. We only need to note that uniformly for all $\mathbf{X}_{e,i}$, $1\leq j\leq p$, and $\boldsymbol{\beta},\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathcal{B}$, there hold

$$
\begin{aligned}
&\left|\left(X_{i,j}-\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)X_j\right)\right)\left(G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)-G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star\right)\right)\right| \\
&\leq CqD_{q,0}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\left\|G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)X_j\right)\right. \\
&\left.-G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_2\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_2\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)X_j\right)\right\| \\
&\leq\left\|\left(G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)-G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\right)\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)X_j\right)\right\| \\
&+\left\|G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\mathbb{E}_{\mathbf{X}_e}\left(\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1\right)-\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_2\right)\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)X_j\right)\right\| \\
&+\left\|G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_2\right)\left(\Gamma_q^{-1}\left(\boldsymbol{\beta}_1\right)-\Gamma_q^{-1}\left(\boldsymbol{\beta}_2\right)\right)\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)X_j\right)\right\| \\
&+\left\|G\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\mathbb{E}_{\mathbf{X}_e}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_2\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}_2\right)\left(\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1\right)-\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_2\right)\right)X_j\right)\right\| \\
&\leq C\sqrt{p}q^2D_{q,0}^3D_{q,1}\left\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right\|.
\end{aligned}
$$

□

Lemma 1.19. *Suppose that Assumption 2.1, Assumption 2.2(i)-(iii), and Assumption 1.6 hold with*

$\upsilon_G \geq 1$, *and that* $\chi_{1,n} \to 0$ *as* $n \to \infty$, *then we have that*

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right) \Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right) \left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right) R_q\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)\right) \right\| = O_p\left(\sqrt{p} q D_{q,0}^2 \mathcal{E}_{q,0}\right),$$

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right) \Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right) \left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right) \varepsilon_j\right) \right\| = O_p\left(\sqrt{p}\chi_{1,n}\right),$$

and

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(R_q\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right) \mathbf{X}_i + \varepsilon_i \mathbf{X}_i\right) \right\| = O_p\left(\sqrt{p}\mathcal{E}_{q,0} + \sqrt{p(\log p)/n}\right).$$

Proof of Lemma 1.19. For the first result, we note that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right) \Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right) \left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right) R_q\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)\right) \right\| \\
&= O_p\left(\sqrt{p} \sup_{\boldsymbol{\beta}\in\mathcal{B},\mathbf{X}_e\in\mathcal{X}_e} \left\| \boldsymbol{r}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\right\| \sup_{\boldsymbol{\beta}\in\mathcal{B},\mathbf{X}_e\in\mathcal{X}_e} \left\| \boldsymbol{r}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right) R_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\right\| \right) \\
&= O_p\left(\sqrt{p} q D_{q,0}^2 \mathcal{E}_{q,0}\right).
\end{aligned}$$

For the second result, we first have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right) \varepsilon_j \right\| = O_p\left(\sqrt{pqD_{q,0}^2 \log\left(pqD_{q,1}n\right)/n} \right),$$

due to the fact that $\left|r_l\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)\varepsilon_j\right| \leq CD_{q,0}$ and $\left\|\left(\partial r_l\left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/\partial\boldsymbol{\beta}\right)\varepsilon_j\right\| \leq C\sqrt{p}D_{q,1}$ for all $0 \leq l \leq q$. So

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right) \Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right) \left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right) \varepsilon_j\right) \right\| \\
&= O_p\left(\sqrt{pq} D_{q,0} \sqrt{pqD_{q,0}^2 \log\left(pqD_{q,1}n\right)/n} \right) = O_p\left(\sqrt{p}\chi_{1,n}\right).
\end{aligned}$$

Finally for the third result, we have that $\left\|\frac{1}{n}\sum_{i=1}^{n} R_q\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right)\mathbf{X}_i\right\| = O_p\left(\sqrt{p}\mathcal{E}_{q,0}\right)$ and $\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_i\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right)$.

Combine the above results, we finish the proof. □

Now we are ready to prove Lemma 1.4 in the main text.

Proof of Lemma 1.4. We note that

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}|\,\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i \\ &= \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\widehat{\boldsymbol{\pi}}_{q,n,k} - \boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\boldsymbol{\pi}_q^{\star}\right)\mathbf{X}_i - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(G\left(z_{i,k}\right) - G\left(\boldsymbol{z}_i^{\star}\right)\right)\mathbf{X}_i \\ &+ \frac{\delta_k}{n}\sum_{i=1}^{n}R_q\left(z_{i,k}\right)\mathbf{X}_i + \frac{\delta}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_i.\end{aligned}$$

Now we look at the $\widehat{\boldsymbol{\pi}}_{q,n,k} - \boldsymbol{\pi}_q^{\star}$. Define $\Gamma_{q,n,k} = \Gamma_{q,n}\left(\boldsymbol{\beta}_k\right)$, we have that

$$\begin{aligned}\widehat{\boldsymbol{\pi}}_{q,n,k} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{i,k}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{i,k}\right)y_i\right) \\ &= \boldsymbol{\pi}_q^{\star} - \Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{i,k}\right)\left(G\left(z_{i,k}\right) - G\left(\boldsymbol{z}_i^{\star}\right)\right)\right) + \Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{i,k}\right)R_q\left(z_{i,k}\right)\right) \\ &+ \Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{i,k}\right)\varepsilon_i\right).\end{aligned}$$

Take the above expression of $\widehat{\boldsymbol{\pi}}_{q,n,k} - \boldsymbol{\pi}_q^{\star}$ into the update of $\boldsymbol{\beta}_k$, we have that

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(z_{i,k},\boldsymbol{\beta}_k\right)\right)\left(G\left(z_{i,k}\right) - G\left(z_i^{\star}\right)\right) \\ &- \frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)R_q\left(z_{j,k}\right) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)\varepsilon_j\right) \\ &+ \frac{\delta_k}{n}\sum_{i=1}^{n}\left(R_q\left(z_{i,k}\right)\mathbf{X}_i + \varepsilon_i\mathbf{X}_i\right).\end{aligned}$$

If we define

$$
\begin{aligned}
\mathfrak{R}_{n,k} = & \mathbb{E}\left(\mathbf{X} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right)\right)\right) \\
& - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}^\star\right)\right)\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}^\star\right)\right)\right) \\
& - \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}^\star\right)\right)\right) \\
& - \frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i \boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)R_q\left(z_{j,k}\right) + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)\varepsilon_j\right) \\
& + \frac{\delta_k}{n}\sum_{i=1}^{n}\left(R_q\left(z_{i,k}\right)\mathbf{X}_i + \varepsilon_i\mathbf{X}_i\right),
\end{aligned}
$$

we have that

$$
\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \delta_k\mathbb{E}\left[\left(\mathbf{X} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right), \boldsymbol{\beta}_k\right)\right)\left(G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}_k\right)\right) - G\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right)\right)\right)\right] + \delta_k\mathfrak{R}_{n,k}.
$$

It remains to verify the order of $\sup_{k\geq 1}\left\|\mathfrak{R}_{n,k}\right\|$, which is done based on Lemma 1.17, Lemma 1.18, and Lemma 1.19. □

Now we prove Lemma 1.5 and Lemma 1.6 in the main text.

Proof of Lemma 1.5. Recall that

$$
\Psi_q\left(t, \boldsymbol{\beta}\right) = \mathbb{E}\left[G'\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star\right) + t\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathfrak{X}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}^{\mathrm{T}}\right)\right].
$$

We have that

$$
\begin{aligned}
& \sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right) - \Psi_q^\star\right\| \\
& \leq \sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right) - \mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\right)\mathbf{X}_i^{\mathrm{T}}\right\| \\
& + \sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right) - \Psi_q\left(t, \boldsymbol{\beta}\right)\right\| \\
& + \sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}_n}\left\|\Psi_q\left(t, \boldsymbol{\beta}\right) - \Psi_q^\star\right\|.
\end{aligned}
$$

From Lemma 1.17, we know that

$$\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)-\mathfrak{X}_{q}\left(z,\boldsymbol{\beta}\right)\right\|=O_{p}\left(\sqrt{p}qD_{q,0}^{2}\chi_{1,n}\right),$$

and as a result,

$$\begin{aligned}&\sup_{0\leq t\leq1,\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_{i}^{\star}+t\mathbf{X}_{i}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)-\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)\mathbf{X}_{i}^{\mathrm{T}}\right\|\\&=O_{p}\left(pqD_{q,0}^{2}\chi_{1,n}\right).\end{aligned}$$

For the second term, we have that

$$\begin{aligned}&\sup_{0\leq t\leq1,\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_{i}^{\star}+t\mathbf{X}_{i}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_{i}\mathbf{X}_{i}^{\mathrm{T}}-\mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_{i}^{\mathrm{T}}\right)-\Psi_{q}\left(t,\boldsymbol{\beta}\right)\right\|\\&=O_{p}\left(\sqrt{p^{3}q^{2}D_{q,0}^{4}\log\left(pqD_{q,0}D_{q,1}n\right)/n}\right)=O_{p}\left(p\chi_{1,n}\right),\end{aligned}$$

due to the fact that

$$\left|G'\left(z_{i}^{\star}+t\mathbf{X}_{i}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(X_{i,s}X_{i,t}-\left(\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)_{s}X_{i,t}\right)\right|\leq CqD_{q.0}^{2},$$

and

$$\begin{aligned}&\left|G'\left(z_{i}^{\star}+t\mathbf{X}_{i}^{\mathrm{T}}\Delta\boldsymbol{\beta}_{1}\right)\left(X_{i,s}X_{i,t}-\left(\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_{1}\right),\boldsymbol{\beta}_{1}\right)\right)_{s}X_{i,t}\right)\right.\\&\left.-G'\left(z_{i}^{\star}+t\mathbf{X}_{i}^{\mathrm{T}}\Delta\boldsymbol{\beta}_{2}\right)\left(X_{i,s}X_{i,t}-\left(\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_{2}\right),\boldsymbol{\beta}_{2}\right)\right)_{s}X_{i,t}\right)\right|\\&\leq C\sqrt{p}q^{2}D_{q,0}^{3}D_{q,1}\left\|\boldsymbol{\beta}_{1}-\boldsymbol{\beta}_{2}\right\|.\end{aligned}$$

Finally,

$$\begin{aligned}&\sup_{0\leq t\leq1,\boldsymbol{\beta}\in\mathcal{B}_{n}}\left\|\Psi_{q}\left(t,\boldsymbol{\beta}\right)-\Psi_{q}^{\star}\right\|\\&\leq\sup_{0\leq t\leq1,\boldsymbol{\beta}\in\mathcal{B}_{n}}\left\|\mathbb{E}\left[G'\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}^{\star}\right)+t\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)-G'\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}^{\star}\right)\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}^{\mathrm{T}}\right)\right]\right\|\\&+\sup_{\boldsymbol{\beta}\in\mathcal{B}_{n}}\left\|\mathbb{E}\left[G'\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}^{\star}\right)\right)\left(\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)-\mathfrak{X}_{q}\left(z\left(\mathbf{X}_{e},\boldsymbol{\beta}^{\star}\right),\boldsymbol{\beta}^{\star}\right)\mathbf{X}^{\mathrm{T}}\right)\right]\right\|.\end{aligned}$$

Obviously the first term is bounded by $C\sqrt{p^{3}}qD_{q,0}^{2}\sup_{\boldsymbol{\beta}\in\mathcal{B}_{n}}\left\|\Delta\boldsymbol{\beta}\right\|$, while the second term is bounded

by

$$
\begin{aligned}
&Cp \sup_{\mathbf{X}_e,\widetilde{\mathbf{X}}_e} \left\| \left(\boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta} \right) \right) \Gamma_q^{-1}\left(\boldsymbol{\beta} \right) \boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right) - \left(\boldsymbol{r}_q^{\mathrm{T}} \left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta}^\star \right) \right) \Gamma_q^{-1}\left(\boldsymbol{\beta}^\star \right) \boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star \right) \right) \right) \right\| \\
&\leq Cp \sup_{\mathbf{X}_e,\widetilde{\mathbf{X}}_e} \left\| \left(\boldsymbol{r}_q\left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta} \right) \right) - \boldsymbol{r}_q\left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta}^\star \right) \right) \right)^{\mathrm{T}} \Gamma_q^{-1}\left(\boldsymbol{\beta} \right) \boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right\| \\
&+ Cp \sup_{\mathbf{X}_e,\widetilde{\mathbf{X}}_e} \left\| \boldsymbol{r}_q\left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta}^\star \right) \right)^{\mathrm{T}} \left(\Gamma_q^{-1}\left(\boldsymbol{\beta} \right) - \Gamma_q^{-1}\left(\boldsymbol{\beta}^\star \right) \right) \boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) \right\| \\
&+ Cp \sup_{\mathbf{X}_e,\widetilde{\mathbf{X}}_e} \left\| \boldsymbol{r}_q\left(z\left(\widetilde{\mathbf{X}}_e, \boldsymbol{\beta}^\star \right) \right)^{\mathrm{T}} \Gamma_q^{-1}\left(\boldsymbol{\beta}^\star \right) \left(\boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta} \right) \right) - \boldsymbol{r}_q\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}^\star \right) \right) \right) \right\| \\
&\leq C\sqrt{p^3} q^2 D_{q,0}^3 D_{q,1} \sup_{\boldsymbol{\beta}\in\mathcal{B}_n} \left\| \Delta\boldsymbol{\beta} \right\|.
\end{aligned}
$$

So

$$
\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}_n} \left\| \Psi_q\left(t, \boldsymbol{\beta} \right) - \Psi_q^\star \right\| = O_p\left(\sqrt{p^3} q^2 D_{q,0}^3 D_{q,1} \sup_{\boldsymbol{\beta}\in\mathcal{B}_n} \left\| \Delta\boldsymbol{\beta} \right\| \right).
$$

Combine the above results, we have that

$$
\begin{aligned}
&\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}_n} \left\| \frac{1}{n} \sum_{i=1}^n G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}} \Delta\boldsymbol{\beta} \right) \left(\mathbf{X}_i \mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) \mathbf{X}_i^{\mathrm{T}} \right) - \Psi_q^\star \right\| \\
&= O_p\left(pqD_{q,0}^2 \chi_{1,n} + \sqrt{p^3} q^2 D_{q,0}^3 D_{q,1} \sup_{\boldsymbol{\beta}\in\mathcal{B}_n} \left\| \Delta\boldsymbol{\beta} \right\| \right).
\end{aligned}
$$

□

Proof of Lemma 1.6. According to Theorem 1.7, we have that $\sup_{k\geq k_{1,n}^{SBGD}+1} \left\| \Delta\boldsymbol{\beta}_k \right\| = O_p\left(\chi_{2,n} \right)$. To prove the lemma, we first show that

$$
\begin{aligned}
&\sup_{k\geq k_{1,n}^{SBGD}+1} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \boldsymbol{r}_{q,i,k}^{\mathrm{T}} \Gamma_{q,n,k}^{-1} \left(\frac{1}{n} \sum_{j=1}^n \boldsymbol{r}_{q,j,k} R_{q,j,k} + \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{q,j,k} \varepsilon_j - \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{q,j}^\star \varepsilon_j \right) \right\| \\
&= O_p\left(\sqrt{p} qD_{q,0}^2 \mathcal{E}_{q,0} + \sqrt{pq} D_{q,0} \chi_{2,n} \chi_{3,n} \right),
\end{aligned}
$$

where $\chi_{3,n} = \sqrt{p^2 q D_{q,1}^2 \log\left(pqD_{q,2} n \right) / n}$. Note that

$$
\begin{aligned}
\sup_{k\geq k_{1,n}^{SBGD}+1} \left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{q,i,k} \varepsilon_i - \frac{1}{n} \sum_{i=1}^n \boldsymbol{r}_{q,i}^\star \varepsilon_i \right\| &= \sup_{k\geq k_{1,n}^{SBGD}+1} \left\| \left\{ \int_0^1 \frac{1}{n} \sum_{i=1}^n \varepsilon_i \boldsymbol{r}_q'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}} \Delta\boldsymbol{\beta}_k \right) \mathbf{X}_i^{\mathrm{T}} dt \right\} \Delta\boldsymbol{\beta}_k \right\| \\
&\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \boldsymbol{r}_q'\left(X_{0,i} + \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right) \mathbf{X}_i^{\mathrm{T}} \right\| \sup_{k\geq k_{1,n}^{SBGD}+1} \left\| \Delta\boldsymbol{\beta}_k \right\|,
\end{aligned}
$$

Obviously, we have that $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \boldsymbol{r}_q'\left(X_{0,j}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right\| = O_p\left(\chi_{3,n}\right)$ due to the fact that $\left|\varepsilon_i r_s'\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)X_t\right| \leq CD_{q,1}$ and $\left\|\partial\varepsilon_i r_s'\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)X_t/\partial\boldsymbol{\beta}\right\| \leq C\sqrt{p}D_{q,2}$, so

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j}^{\star}\varepsilon_j\right\| = O_p\left(\chi_{2,n}\chi_{3,n}\right),$$

which leads to the result if we further note that

$$\begin{aligned}
&\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}R_{q,j,k}+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j}^{\star}\varepsilon_j\right)\right\| \\
&= O_p\left(\sqrt{pq}D_{q,0}\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}R_{q,j,k}+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j}^{\star}\varepsilon_j\right\|\right) \\
&= O_p\left(\sqrt{p}qD_{q,0}^2\mathcal{E}_{q,0}+\sqrt{pq}D_{q,0}\chi_{2,n}\chi_{3,n}\right).
\end{aligned}$$

Next we show that

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}-\mathbb{E}\left(\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)\right\| = O_p\left(p\sqrt{q^3}D_{q,0}^2D_{q,1}\chi_{2,n}\right).$$

$$\begin{aligned}
&\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| \\
&\leq \sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)\Gamma_{q,n,k}^{-1}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_{q,n,k}^{-1}\right\| \\
&+ \sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_{q,n,k}^{-1}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\|.
\end{aligned}$$

The first term is obviously bounded in probability by

$$\begin{aligned}
&C\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\right)-\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}\right\| \\
&\leq Cp\sqrt{q}D_{q,1}\left\|\boldsymbol{\beta}_k-\boldsymbol{\beta}^{\star}\right\| = Cp\sqrt{q}D_{q,1}\chi_{2,n}.
\end{aligned}$$

The second term is bounded by

$$\begin{aligned}
&\sup_{k\geq k_{1,n}^{*}}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\right\|\sup_{k\geq k_{1,n}^{*}}\left\|\Gamma_{q,n,k}^{-1}-\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| \\
&\leq C\sqrt{pq}D_{q,0}\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Gamma_{q,n,k}^{-1}-\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\|.
\end{aligned}$$

Now we provide an upper bound for $\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Gamma_{q,n,k}^{-1}-\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\|$. Note that

$$\begin{aligned}
\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Gamma_{q,n,k}^{-1}-\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| &= O_p\left(\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Gamma_{q,n,k}-\Gamma_q\left(\boldsymbol{\beta}^{\star}\right)\right\|\right) \\
&= O_p\left(\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\Gamma_{q,n,k}-\Gamma_{q,n}\left(\boldsymbol{\beta}^{\star}\right)\right\|+\left\|\Gamma_{q,n}\left(\boldsymbol{\beta}^{\star}\right)-\Gamma_q\left(\boldsymbol{\beta}^{\star}\right)\right\|\right) \\
&= O_p\left(\sqrt{p}qD_{q,0}D_{q,1}\chi_{2,n}+\chi_{1,n}\right)=O_p\left(\sqrt{p}qD_{q,0}D_{q,1}\chi_{2,n}\right).
\end{aligned}$$

So

$$\begin{aligned}
&\sup_{k\geq k_{1,n}^{*}}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_{q,n,k}^{-1}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| \\
&=O_p\left(p\sqrt{q^3}D_{q,0}^2D_{q,1}\chi_{2,n}\right),
\end{aligned}$$

and together

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\|=O_p\left(p\sqrt{q^3}D_{q,0}^2D_{q,1}\chi_{2,n}\right).$$

Moreover, note that $\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)-\mathbb{E}\left(\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)\right\|=O_p\left(\sqrt{p^3D_{q,0}^2\log\left(pn\right)/n}\right)$, so we have shown the results.

Based on the above results, we have that

$$
\begin{aligned}
&\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}R_{q,j,k}+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j\right)+\frac{1}{n}\sum_{i=1}^{n}R_q\left(z_{i,k}\right)\mathbf{X}_i-\frac{1}{n}\sum_{i=1}^{n}\mathfrak{X}\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)\varepsilon_j\right\| \\
&\leq \sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_{q,j,k}R_{q,j,k}+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j,k}\varepsilon_j-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j}^{\star}\varepsilon_j\right)\right\| \\
&+\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_{q,i,k}^{\mathrm{T}}\Gamma_{q,n,k}^{-1}-\mathbb{E}\left(\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}\right)\Gamma_q^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_{q,j}^{\star}\varepsilon_j\right)\right\| \\
&+\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\frac{1}{n}\sum_{i=1}^{n}R_q\left(z_{i,k}\right)\mathbf{X}_i\right\| \\
&=O_p\left(\sqrt{p}qD_{q,0}^2\mathcal{E}_{q,0}+\sqrt{pq}D_{q,0}\chi_{2,n}\chi_{3,n}+p\sqrt{q^3}D_{q,0}^2D_{q,1}\chi_{2,n}\sqrt{\left(qD_{q,0}^2\log q\right)/n}\right)
\end{aligned}
$$

□

1.7.2 Proofs of Theorems

Proof of Theorem 1.1

Proof. We first prove Theorem 1.1(i). Recall that $\Delta\boldsymbol{\beta}_{e,k}=\boldsymbol{\beta}_{e,k}-\boldsymbol{\beta}_e^{\star}$ and $\varepsilon_i=y_i-G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right)$. We have that

$$
\Delta\boldsymbol{\beta}_{e,k+1}=\Delta\boldsymbol{\beta}_{e,k}-\frac{\delta}{n}\sum_{i=1}^{n}\left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right)-G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right)-\varepsilon_i\right)\mathbf{X}_{e,i},
$$

so

$$
\left\|\Delta\boldsymbol{\beta}_{e,k+1}\right\|\leq\left\|\Delta\boldsymbol{\beta}_{e,k}-\frac{\delta}{n}\sum_{i=1}^{n}\left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right)-G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right)\right)\mathbf{X}_{e,i}\right\|+\left\|\frac{\delta}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\|.
$$

Note that mean value theorem leads to

$$
\begin{aligned}
&\Delta\boldsymbol{\beta}_{e,k}-\frac{\delta}{n}\sum_{i=1}^{n}\left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right)-G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right)\right)\mathbf{X}_{e,i} \\
&=\Delta\boldsymbol{\beta}_{e,k}-\int_0^1\left\{\frac{\delta}{n}\sum_{i=1}^{n}G'\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}+t\mathbf{X}_{e,i}^{\mathrm{T}}\Delta\boldsymbol{\beta}_{e,k}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\Delta\boldsymbol{\beta}_{e,k}\right\}dt \\
&=\int_0^1\left\{\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e^{\star}+t\Delta\boldsymbol{\beta}_{e,k}\right)\right)\Delta\boldsymbol{\beta}_{e,k}\right\}dt,
\end{aligned}
$$

where the integration is understood to be element-wise, and $\boldsymbol{\beta}_e^{\star}+t\Delta\boldsymbol{\beta}_{e,k}\in\mathcal{B}_e$ due to convexity of $\mathcal{B}_e$.

We next provide a uniform upper bound for $\overline{\lambda}\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e\right)\right)$ and lower bound for $\underline{\lambda}\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e\right)\right)$ with respect to $\boldsymbol{\beta}_e \in \mathcal{B}_e$ in probability. Since Assumption 2.2 holds, we have that $G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}\right) X_{i,t}X_{i,s}$ is bounded by $\|G\|_\infty$ and $\left\|\partial G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}\right) X_{i,t}X_{i,s}/\partial\boldsymbol{\beta}\right\| \leq C\sqrt{p}$. Then according to Lemma 1.7, we have that

$$\sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\left\|M_n\left(\boldsymbol{\beta}_e\right)-M\left(\boldsymbol{\beta}_e\right)\right\| = O_p\left(\sqrt{\frac{p^3\log n}{n}}\right).$$

Since $p^5(\log p)^2 n^{-1}\to 0$ holds, $\sqrt{p^3\left(\log n\right)/n}\to 0$ holds, so

$$\sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\left|\overline{\lambda}\left(M_n\left(\boldsymbol{\beta}_e\right)\right)-\overline{\lambda}\left(M\left(\boldsymbol{\beta}_e\right)\right)\right| = o_p\left(1\right),$$

and

$$\sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\left|\underline{\lambda}\left(M_n\left(\boldsymbol{\beta}_e\right)\right)-\underline{\lambda}\left(M\left(\boldsymbol{\beta}_e\right)\right)\right| = o_p\left(1\right).$$

Due to Assumption 2.2(iv), with probability going to 1, there holds,

$$\underline{\lambda}_e/2 \leq \inf_{\boldsymbol{\beta}_e\in\mathcal{B}}\underline{\lambda}\left(M_n\left(\boldsymbol{\beta}_e\right)\right) \leq \sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\overline{\lambda}\left(M_n\left(\boldsymbol{\beta}_e\right)\right) \leq 3\overline{\lambda}_e/2.$$

Since $\delta < 2/(3\overline{\lambda}_e)$, we have that with probability going to 1, there holds

$$0 \leq \inf_{\boldsymbol{\beta}_e\in\mathcal{B}}\overline{\lambda}\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e\right)\right) \leq \sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\overline{\lambda}\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e\right)\right) \leq 1-\underline{\lambda}_e\delta/2.$$

Based on the above inequality, we have that with probability going to 1, there holds

$$\begin{aligned}
&\left\|\int_0^1\left\{\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e^\star+t\Delta\boldsymbol{\beta}_{e,k}\right)\right)\Delta\boldsymbol{\beta}_{e,k}\right\}dt\right\| \\
&\leq \int_0^1\left\{\sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\overline{\lambda}\left(I_{p+1}-\delta M_n\left(\boldsymbol{\beta}_e\right)\right)\right\}dt\cdot\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| \leq \left(1-\underline{\lambda}_e\delta/2\right)\cdot\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|.
\end{aligned}$$

So with probability going to 1, for all k there holds

$$\begin{aligned}
\left\|\Delta\boldsymbol{\beta}_{e,k+1}\right\| &\leq \left(1-\underline{\lambda}_e\delta/2\right)\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|+\delta\left\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbf{X}_{e,i}\right\| \\
&\leq\cdots\leq\left(1-\underline{\lambda}_e\delta/2\right)^k\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|+\delta\sum_{j=1}^k\left(1-\underline{\lambda}_e\delta/2\right)^{j-1}\left\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbf{X}_{e,i}\right\| \\
&\leq\left(1-\underline{\lambda}_e\delta/2\right)^k\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|+2\underline{\lambda}_e^{-1}\left\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbf{X}_{e,i}\right\|.
\end{aligned}$$

Note that for any $\tau > 0$,

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{X}_{e,i}\right\| > \tau\right) \leq \sum_{j=0}^{p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{e,j,i}\right| > \frac{\tau}{\sqrt{p+1}}\right)$$
$$\leq \sum_{j=0}^{p} 2\exp\left(Cn\tau^2/p\right) = 2\exp\left(C_1 \log p - C_2 n\tau^2/p\right),$$

so

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \mathbf{X}_{e,i}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right).$$

Then for k such that

$$\left(1 - \underline{\lambda}_e \delta/2\right)^k \left\|\Delta\boldsymbol{\beta}_{e,1}\right\| \leq \sqrt{p\left(\log p\right)/n},$$

or equivalently,

$$k \geq k_{1,n}^{BGD} = \frac{\log\left\|\Delta\boldsymbol{\beta}_{e,1}\right\| + \frac{1}{2}\log\left(n/\left(p\log p\right)\right)}{-\log\left(1 - \underline{\lambda}_e\delta/2\right)},$$

we have that

$$\left\|\Delta\boldsymbol{\beta}_{e,k+1}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right).$$

This proves Theorem 1.1(i).

Next we prove Theorem 1.1(ii). For any $k \geq k_{1,n}^{BGD} + 1$, there holds

$$\Delta\boldsymbol{\beta}_{e,k+1} = \Delta\boldsymbol{\beta}_{e,k} - \frac{\delta}{n}\sum_{i=1}^{n}\left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right) - G\left(\boldsymbol{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_e^{\star}\right) - \varepsilon_i\right)\mathbf{X}_{e,i},$$
$$= \left(I_{p+1} - \delta M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right)\right)\Delta\boldsymbol{\beta}_{e,k} + \frac{\delta}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i},$$

where $\overline{\boldsymbol{\beta}}_{e,k}$ is element-wise and lies between $\boldsymbol{\beta}_{e,k}$ and $\boldsymbol{\beta}_e^{\star}$. Since $\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right)$ for $k \geq k_{1,n}^{BGD} + 1$, $\left\|\Delta\overline{\boldsymbol{\beta}}_{e,k}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right)$ also holds. Note that

$$\left\|M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right) - M\left(\boldsymbol{\beta}_e^{\star}\right)\right\| \leq \left\|M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right) - M_n\left(\boldsymbol{\beta}_e^{\star}\right)\right\| + \left\|M_n\left(\boldsymbol{\beta}_e^{\star}\right) - M\left(\boldsymbol{\beta}_e^{\star}\right)\right\|.$$

For the second term, $\left\|M_n\left(\boldsymbol{\beta}_e^{\star}\right) - M\left(\boldsymbol{\beta}_e^{\star}\right)\right\| = O_p\left(\sqrt{p^2\left(\log p\right)/n}\right)$ obviously holds. For the first term,

since G is twice differentiable with bounded derivatives, we have that

$$\begin{aligned}\sup_{k\geq k_{1,n}^{BGD}+1}\left\|M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right)-M_n\left(\boldsymbol{\beta}_e^{\star}\right)\right\| &\leq \sup_{k\geq k_{1,n}^{BGD}+1}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right\|\left|G''\left(\mathbf{X}_{e,i}^{\mathrm{T}}\check{\boldsymbol{\beta}}_{e,k}\right)\right|\left|\mathbf{X}_{e,i}^{\mathrm{T}}\Delta\overline{\boldsymbol{\beta}}_{e,k}\right|.\\ &\leq C\sqrt{p^3}\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\overline{\boldsymbol{\beta}}_{e,k}-\boldsymbol{\beta}_e^{\star}\right\|=O_p\left(\sqrt{p^4\left(\log p\right)/n}\right),\end{aligned}$$

where $\check{\boldsymbol{\beta}}_{e,k}$ lies somewhere between $\overline{\boldsymbol{\beta}}_{e,k}$ and $\boldsymbol{\beta}_e^{\star}$ and is also element-wise, and the second last inequality comes from the fact that $\left\|\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right\|\leq p$ and $\left|\mathbf{X}_{e,i}^{\mathrm{T}}\Delta\overline{\boldsymbol{\beta}}_{e,k}\right|\leq\left\|\mathbf{X}_{e,i}\right\|\left\|\Delta\overline{\boldsymbol{\beta}}_{e,k}\right\|$. This implies that

$$\sup_{k\geq k_{1,n}+1}\left\|M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right\|=O_p\left(\sqrt{p^4\left(\log p\right)/n}\right).$$

Define $\omega_k=\left(M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\Delta\boldsymbol{\beta}_{e,k}$. Obviously, there holds

$$\begin{aligned}\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\omega_k\right\| &\leq \left(\sup_{k\geq k_{1,n}^{BGD}+1}\left\|M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right\|\right)\left(\sup_{k\geq k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|\right)\\ &=O_p\left(\sqrt{p^5\left(\log p\right)^2/n^2}\right),\end{aligned}$$

which is $o_p\left(n^{-1/2}\right)$ according to Assumption 2.

Based on the above result, we have that for any $k\geq 1$,

$$\begin{aligned}\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}+1} &= \left(I_{p+1}-\delta M_n\left(\overline{\boldsymbol{\beta}}_{e,k+k_{1,n}^{BGD}}\right)\right)\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}}-\frac{\delta}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\\ &=\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}}-\delta\omega_{k+k_{1,n}^{BGD}}-\frac{\delta}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\\ &=\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^k\Delta\boldsymbol{\beta}_{e,k_{1,n}^{BGD}+1}-\delta\sum_{j=0}^{k-1}\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^j\omega_{k+k_{1,n}^{BGD}-j}\\ &\quad-\delta\left(\sum_{j=0}^{k-1}\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^j\right)\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right).\end{aligned}$$

For the first part on the RHS of the last equality, we have that

$$\begin{aligned}\left\|\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^k\Delta\boldsymbol{\beta}_{e,k_{1,n}^{BGD}+1}\right\| &\leq\left(1-\underline{\lambda}_e\delta\right)^k\left\|\Delta\boldsymbol{\beta}_{e,k_{1,n}^{BGD}+1}\right\|\\ &=\left(1-\underline{\lambda}_e\delta\right)^k O_p\left(\sqrt{p\left(\log p\right)/n}\right).\end{aligned}$$

For the second part, we have that

$$\begin{aligned}\left\|\delta\sum_{j=0}^{k-1}\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^j\omega_{k+k_{1,n}^{BGD}-j}\right\| &\leq \delta\sum_{j=0}^{\infty}\left(1-\underline{\lambda}_e\delta\right)^j\left\|\omega_{k+k_{1,n}^{BGD}-j}\right\| \\ &\leq \underline{\lambda}_e^{-1}\sup_{k\geq 1}\left\|\omega_{k+k_{1,n}^{BGD}}\right\| = O_p\left(\sqrt{p^5\left(\log p\right)^2/n^2}\right) \\ &= o_p\left(n^{-1/2}\right).\end{aligned}$$

For the third part, we have that

$$\begin{aligned}&\left\|\left(\sum_{j=0}^{k-1}\delta\left(I_{p+1}-\delta M\left(\boldsymbol{\beta}_e^{\star}\right)\right)^j\right)\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right)-M_n^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right)\right\| \\ &\leq \sum_{j=k}^{\infty}\delta\left(1-\underline{\lambda}_e\delta\right)^j\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| = \left(1-\underline{\lambda}_e\delta\right)^k\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| \\ &= \left(1-\underline{\lambda}_e\delta\right)^k O_p\left(\sqrt{p\left(\log p\right)/n}\right).\end{aligned}$$

This implies that when $\left(1-\underline{\lambda}_e\delta\right)^{k_{2,n}^{BGD}}\sqrt{p\log p}\to 0$, we have that

$$\sup_{k\geq k_{2,n}^{BGD}+1}\left\|\sqrt{n}\Delta\boldsymbol{\beta}_{e,k+k_{1,n}^{BGD}}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| = o_p\left(1\right).$$

This proves Theorem 1.1(ii)

Now we prove Theorem 1.1(iii). We first note that for any square matrices A, B, and C, there hold $\|AB\|\leq\overline{\sigma}\left(A\right)\|B\|$ and $\|ABC\|\leq\overline{\sigma}\left(A\right)\|BC\|\leq\overline{\sigma}\left(A\right)\overline{\sigma}\left(B\right)\|C\|$. So

$$\begin{aligned}\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| &= \left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\left(M_n\left(\widehat{\boldsymbol{\beta}}_e\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right)M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| \\ &\leq \overline{\sigma}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)\cdot\overline{\sigma}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\cdot\left\|M_n\left(\widehat{\boldsymbol{\beta}}_e\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right\|,\end{aligned}$$

due to the fact that $M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)$ and $M_n\left(\widehat{\boldsymbol{\beta}}_e\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)$ are both symmetric. Due to Assumption 2.2(iv), we have that $\overline{\sigma}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)=\overline{\lambda}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)\leq\underline{\lambda}_e^{-1}$. Since $\left\|M_n\left(\widehat{\boldsymbol{\beta}}_e\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right\|=o_p\left(1\right)$ holds according to the previous proof, we have that with probability going to 1, $\overline{\sigma}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)=\overline{\lambda}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\leq 2\underline{\lambda}_e^{-1}$. Then with probability going to 1, we have that

$$\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\|\leq 2\underline{\lambda}_e^{-2}\left\|M_n\left(\widehat{\boldsymbol{\beta}}_e\right)-M\left(\boldsymbol{\beta}_e^{\star}\right)\right\| = O_p\left(\sqrt{p^4\left(\log p\right)/n}\right).$$

On the other side, we have that

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right\| \\
&\leq\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\frac{1}{n}\sum_{i=1}^{n}G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right\| \\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right\| \\
&\leq C\sqrt{p^3}\left\|\widehat{\boldsymbol{\beta}}_e-\boldsymbol{\beta}_e^{\star}\right\|+O_p\left(\sqrt{p^2\left(\log p\right)/n}\right)=O_p\left(\sqrt{p^4(\log p)/n}\right).
\end{aligned}$$

Together, we have that

$$\begin{aligned}
\left\|\widehat{\Sigma}_1-\Sigma_1^{\star}\right\| &\leq\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\right\| \\
&+\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right)M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| \\
&+\left\|\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| \\
&\leq\overline{\lambda}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)\overline{\lambda}\left(\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right)\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| \\
&+\overline{\lambda}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)\overline{\lambda}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right\| \\
&+\overline{\lambda}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\overline{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\|.
\end{aligned}$$

Note that $\overline{\lambda}\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right)\leq\underline{\lambda}_e^{-1}$, $\overline{\lambda}\left(\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right)\leq\frac{1}{4}\overline{\lambda}\left(\mathbb{E}\left[\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right)\leq C$, $\overline{\lambda}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right)\leq 2\underline{\lambda}_e^{-1}$ with probability going to 1, and $\overline{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)\leq C$ with probability going to 1, we have that

$$\begin{aligned}
\left\|\widehat{\Sigma}_1-\Sigma_1^{\star}\right\| &\leq C\left\|M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right\| \\
&+C\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}-\mathbb{E}\left[G_i^{\star}\left(1-G_i^{\star}\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right\| \\
&=O_p\left(\sqrt{p^4\left(\log p\right)/n}\right)=o_p\left(1\right),
\end{aligned}$$

which validates the result.

To prove (iv), we only need to show that $\widehat{\sigma}^2(\rho) - \sigma^2(\rho) = o_p(1)$. Note that

$$\left|\widehat{\sigma}_n^2(\rho) - \sigma^2(\rho)\right| = \left|\rho^{\mathrm{T}}\left(\widehat{\Sigma}_1 - \Sigma_1^\star\right)\rho\right| \leq \|\rho\| \left\|\left(\widehat{\Sigma}_1 - \Sigma_1^\star\right)\rho\right\| \leq \|\rho\|^2 \left\|\widehat{\Sigma}_1 - \Sigma_1^\star\right\| \to_p 0$$

given that $\|\rho\| < \infty$ for all n, which validates the result. □

Proof of Theorem 1.2

Proof. We first show Theorem 1.2(i). Note that from the proof in Theorem 1.1, we know that with probability going to 1, we have that

$$\begin{aligned}
\left\|\Delta\boldsymbol{\beta}_{e,k+1}\right\| &\leq \sup_{\boldsymbol{\beta}_e \in \mathcal{B}} \overline{\lambda}\left(I_{p+1} - \delta_k M_n\left(\boldsymbol{\beta}_e\right)\right)\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| + \delta_k \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{X}_{e,i}\right\| \\
&\leq \left(1 - \underline{\lambda}_e \delta_k/2\right)\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| + \delta_k \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{X}_{e,i}\right\| \leq \cdots \\
&\leq \left(\prod_{j=1}^k \left(1 - \underline{\lambda}_e \delta_j/2\right)\right)\left\|\Delta\boldsymbol{\beta}_{e,1}\right\| + \left\{\sum_{j=0}^{k-1} \delta_{k-j}\left(\prod_{l=0}^{j-1}\left(1 - \underline{\lambda}_e \delta_{k-l}/2\right)\right)\right\}\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbf{X}_{e,i}\right\|,
\end{aligned} \tag{1.11}$$

where $\prod_{l=0}^{j-1}\left(1 - \underline{\lambda}_e \delta_{k-l}/2\right) = 1$ if $j = 0$.

For the first term on the RHS of (1.11), since $e^x \geq 1 + x$ for all x, we have $1 - \underline{\lambda}_e \delta_j/2 \leq \exp\left(-\underline{\lambda}_e \delta_j/2\right)$ for all j. Define $S_0 = 0$ and $S_j = \sum_{l=1}^j \delta_l$ for $j \geq 1$, we have that

$$\left(\prod_{j=1}^k \left(1 - \underline{\lambda}_e \delta_j/2\right)\right)\left\|\Delta\boldsymbol{\beta}_{e,1}\right\| \leq \exp\left(-\frac{\underline{\lambda}_e}{2}\sum_{j=1}^k \delta_j\right)\left\|\Delta\boldsymbol{\beta}_{e,1}\right\| = \exp\left(-\frac{\underline{\lambda}_e S_k}{2}\right)\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|.$$

Next we show that $\sum_{j=0}^{k-1} \delta_{k-j}\left(\prod_{l=0}^{j-1}\left(1 - \underline{\lambda}_e \delta_{k-l}/2\right)\right)$ is upper bounded by $\exp\left(\underline{\lambda}_e \delta_{k+1}/2\right)$ up to

some constant scale that is independent of k. Since $\limsup_k \delta_{k-1}/\delta_k < \infty$, we have that

$$\begin{aligned}
&\sum_{j=0}^{k-1} \delta_{k-j} \left(\prod_{l=0}^{j-1} (1 - \underline{\lambda}_e \delta_{k-l}/2) \right) \leq \sum_{j=0}^{k-1} \delta_{k-j} \exp\left(-\frac{\underline{\lambda}_e}{2} \sum_{l=0}^{j-1} \delta_{k-l} \right) \\
&\leq C \sum_{j=0}^{k-1} \delta_{k-j+1} \exp\left(-\frac{\underline{\lambda}_e \left(S_k - S_{k-j} \right)}{2} \right) \\
&= C \exp\left(-\frac{\underline{\lambda}_e S_k}{2} \right) \sum_{j=0}^{k-1} \left(S_{k-j+1} - S_{k-j} \right) \exp\left(\frac{\underline{\lambda}_e S_{k-j}}{2} \right) \\
&\leq 2C \underline{\lambda}_e^{-1} \exp\left(-\frac{\underline{\lambda}_e S_k}{2} \right) \sum_{j=0}^{k-1} \left\{ \exp\left(\frac{\underline{\lambda}_e S_{k-j+1}}{2} \right) - \exp\left(\frac{\underline{\lambda}_e S_{k-j}}{2} \right) \right\} \leq C \exp\left(\frac{\underline{\lambda}_e \delta_{k+1}}{2} \right).
\end{aligned}$$

Then we have that

$$\left\| \Delta \boldsymbol{\beta}_{e,k+1} \right\| = O_p\left(\exp\left(-\frac{\underline{\lambda}_e S_k}{2} \right) \left\| \Delta \boldsymbol{\beta}_{e,1} \right\| \right) + O_p\left(\exp\left(\frac{\underline{\lambda}_e \delta_{k+1}}{2} \right) \sqrt{p \left(\log p \right) / n} \right).$$

When $k \geq \widetilde{k}_{1,n}^{BGD} + 1$, we have that

$$\exp\left(-\frac{\underline{\lambda}_e S_k}{2} \right) \left\| \Delta \boldsymbol{\beta}_{e,1} \right\| \leq \sqrt{p \left(\log p \right) / n},$$

and

$$\exp\left(\frac{\underline{\lambda}_e \delta_{k+1}}{2} \right) \leq e,$$

so $\left\| \Delta \boldsymbol{\beta}_{e,k+1} \right\| = O_p\left(\sqrt{p \left(\log p \right) / n} \right)$. This validates Theorem 1.2(i).

For Theorem 1.2(ii), we know that for $k \geq \widetilde{k}_{1,n}^{BGD} + 1$, $\left\| \Delta \boldsymbol{\beta}_{e,k} \right\| = O_p\left(\sqrt{p \left(\log p \right) / n} \right)$ holds, so we have that

$$\begin{aligned}
\Delta \boldsymbol{\beta}_{e,k+1} &= \left(I_{p+1} - \delta_k M_n\left(\overline{\boldsymbol{\beta}}_{e,k} \right) \right) \Delta \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_{e,i} \\
&= \left(I_{p+1} - \delta_k M\left(\boldsymbol{\beta}_e^{\star} \right) \right) \Delta \boldsymbol{\beta}_{e,k} - \delta_k \left(M_n\left(\overline{\boldsymbol{\beta}}_{e,k} \right) - M\left(\boldsymbol{\beta}_e^{\star} \right) \right) \Delta \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_{e,i},
\end{aligned}$$

where $\overline{\boldsymbol{\beta}}_{e,k}$ lies between $\boldsymbol{\beta}_{e,k}$ and $\boldsymbol{\beta}_e^{\star}$ and is element-wise. Following the proof of Theorem 1.1, we can easily show that

$$\sup_{k \geq \widetilde{k}_{1,n}^{BGD} + 1} \left\| M_n\left(\overline{\boldsymbol{\beta}}_{e,k} \right) - M\left(\boldsymbol{\beta}_e^{\star} \right) \right\| = O_p\left(\sqrt{p^4 \left(\log p \right) / n} \right).$$

Recall that $\omega_k = \left(M_n\left(\overline{\boldsymbol{\beta}}_{e,k}\right) - M\left(\boldsymbol{\beta}_e^\star\right)\right)\Delta\boldsymbol{\beta}_{e,k}$, so

$$\sup_{k\geq \widetilde{k}_{1,n}^{BGD}+1}\|\omega_k\| = O_p\left(\sqrt{p^5\left(\log p\right)^2/n^2}\right) = o_p\left(n^{-1/2}\right).$$

We have that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{e,k+\widetilde{k}_{1,n}^{BGD}+1} &= \left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k}M\left(\boldsymbol{\beta}_e^\star\right)\right)\Delta\boldsymbol{\beta}_{e,k+\widetilde{k}_{1,n}^{BGD}} - \delta_{\widetilde{k}_{1,n}^{BGD}+k}\omega_{\widetilde{k}_{1,n}^{BGD}+k} - \delta_{\widetilde{k}_{1,n}^{BGD}+k}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i} \\
&= \prod_{j=0}^{k-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-j}M\left(\boldsymbol{\beta}_e^\star\right)\right)\Delta\boldsymbol{\beta}_{e,\widetilde{k}_{1,n}^{BGD}+1} \\
&- \sum_{j=0}^{k-1}\left\{\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\prod_{l=0}^{j-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-l}M\left(\boldsymbol{\beta}_e^\star\right)\right)\right\}\omega_{\widetilde{k}_{1,n}^{BGD}+k-j} \\
&- \sum_{j=0}^{k-1}\left\{\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\prod_{l=0}^{j-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-l}M\left(\boldsymbol{\beta}_e^\star\right)\right)\right\}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i},
\end{aligned}$$

where $\prod_{l=0}^{j-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-l}M\left(\boldsymbol{\beta}_e^\star\right)\right) = 1$ if $j=0$. For the first part, define $S_{\widetilde{k}_{1,n}^{BGD},k} = \sum_{j=\widetilde{k}_{1,n}^{BGD}+1}^{\widetilde{k}_{1,n}^{BGD}+k}\delta_j$, we have that

$$\begin{aligned}
\left\|\prod_{j=0}^{k-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-j}M\left(\boldsymbol{\beta}_e^\star\right)\right)\Delta\boldsymbol{\beta}_{e,\widetilde{k}_{1,n}^{BGD}+1}\right\| &\leq \prod_{j=0}^{k-1}\left(1 - \underline{\lambda}_e\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}/2\right)\left\|\Delta\boldsymbol{\beta}_{e,\widetilde{k}_{1,n}^{BGD}+1}\right\| \\
&\leq \exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD},k}/2\right)\left\|\Delta\boldsymbol{\beta}_{e,\widetilde{k}_{1,n}^{BGD}+1}\right\| \\
&= O_p\left(\exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD},k}/2\right)\sqrt{p\left(\log p\right)/n}\right).
\end{aligned}$$

For the second term, we have that

$$\begin{aligned}
&\left\|\sum_{j=0}^{k-1}\left\{\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\prod_{l=0}^{j-1}\left(I_{p+1} - \delta_{\widetilde{k}_{1,n}^{BGD}+k-l}M\left(\boldsymbol{\beta}_e^\star\right)\right)\right\}\omega_{\widetilde{k}_{1,n}^{BGD}+k-j}\right\| \\
&\leq \left\{\sum_{j=0}^{k-1}\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\prod_{l=0}^{j-1}\left(1 - \underline{\lambda}_e\delta_{\widetilde{k}_{1,n}^{BGD}+k-l}/2\right)\right\}\left\{\sup_{k\geq 1}\left\|\omega_{\widetilde{k}_{1,n}^{BGD}+k}\right\|\right\} \\
&\leq \exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD}+k}/2\right)\left\{\sum_{j=0}^{k-1}\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\exp\left(\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD}+k-j}/2\right)\right\}\left\{\sup_{k\geq 1}\left\|\omega_{\widetilde{k}_{1,n}^{BGD}+k}\right\|\right\} \\
&= O_p\left(\sqrt{p^5\left(\log p\right)^2/n^2}\right)
\end{aligned}$$

according to the proof of Theorem 1.2(i). Now we look at the last term. Note that

$$\begin{aligned}\mathcal{M}_{k,n} \equiv: & \sum_{j=0}^{k-1}\left\{\delta_{\widetilde{k}_{1,n}^{BGD}+k-j}\prod_{l=0}^{j-1}\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k-l}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\right\} \\ & = \delta_{\widetilde{k}_{1,n}^{BGD}+k}I_{p+1}+\delta_{\widetilde{k}_{1,n}^{BGD}+k-1}\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)+\cdots \\ & +\delta_{\widetilde{k}_{1,n}^{BGD}+1}\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k-1}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\cdots\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+2}M\left(\boldsymbol{\beta}_e^{\star}\right)\right),\end{aligned}$$

so

$$\mathcal{M}_{k+1,n}=\delta_{\widetilde{k}_{1,n}^{BGD}+k+1}I_{p+1}+\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\mathcal{M}_{k,n}.$$

Note that

$$\begin{aligned}&\mathcal{M}_{k+1,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right) \\ &=\mathcal{M}_{k,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)+\delta_{\widetilde{k}_{1,n}^{BGD}+k+1}M\left(\boldsymbol{\beta}_e^{\star}\right)\left(M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)-\mathcal{M}_{k,n}\right) \\ &=\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\left(\mathcal{M}_{k,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right),\end{aligned}$$

so

$$\begin{aligned}\left\|\mathcal{M}_{k+1,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right\| &\leq \overline{\lambda}\left(I_{p+1}-\delta_{\widetilde{k}_{1,n}^{BGD}+k}M_n\left(\boldsymbol{\beta}_e^{\star}\right)\right)\left\|\mathcal{M}_{k,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right\| \\ &\leq\left(1-\delta_{\widetilde{k}_{1,n}^{BGD}+k}\underline{\lambda}_e\right)\left\|\mathcal{M}_{k,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right\| \\ &\leq\exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD},k}\right)\left\|\mathcal{M}_{1,n}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\right\|.\end{aligned}$$

Then

$$\begin{aligned}&\sum_{j=0}^{k-1}\left\{\delta_{\widetilde{k}_{1,n}^{BGD}+k-1-j}\prod_{l=0}^{j-1}\left(I-\delta_{\widetilde{k}_{1,n}^{BGD}+k-1}M\left(\boldsymbol{\beta}_e^{\star}\right)\right)\right\}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i} \\ &=M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}+O_p\left(\exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD},k}\right)\sqrt{p\left(\log p\right)/n}\right).\end{aligned}$$

So we have

$$\begin{aligned}\left\|\sqrt{n}\Delta\boldsymbol{\beta}_{e,k+\widetilde{k}_{1,n}^{BGD}}-M^{-1}\left(\boldsymbol{\beta}_e^{\star}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| &= O_p\left(\exp\left(-\underline{\lambda}_e S_{\widetilde{k}_{1,n}^{BGD},k}/2\right)\sqrt{p\left(\log p\right)/n}\right) \\ &\quad+O_p\left(\sqrt{p^5\left(\log p\right)^2/n^2}\right) \\ &\quad+O_p\left(\exp\left(-S_{\widetilde{k}_{1,n}^{BGD},k}\right)\sqrt{p\left(\log p\right)/n}\right).\end{aligned}$$

According to the definition of $\widetilde{k}_{2,n}^{BGD}$, we have that for $k \geq \widetilde{k}_{2,n}^{BGD}$, there holds $S_{\widetilde{k}_{1,n}^{BGD},k}/\log p \to \infty$, this proves Theorem 1.2(ii).

The proof of Theorem 1.2(iii) and Theorem 1.2(iv) is the same as that in the proof of Theorem 1.1, so is left out. □

Proof of Theorem 1.3

Proof. Define

$$\eta_{1,n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)\mathbf{X}_i - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right],$$

$$\eta_{2,n} = \left(\frac{1}{n}\sum_{i=1}^{n} G\left(z_i^{\star}\right)\mathbf{X}_i - \mathbb{E}\left[G\left(z_i^{\star}\right)\mathbf{X}_i\right]\right) + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \cdot \mathbf{X}_i.$$

Note that when $\boldsymbol{\beta}^{\star} \in \mathcal{B}$ and $\boldsymbol{\beta}_k \in \mathcal{B}$, we have that $\boldsymbol{\beta}^{\star} + t\Delta\boldsymbol{\beta}_k \in \mathcal{B}$ for all $0 \leq t \leq 1$, so

$$\begin{aligned}\left\|\Delta\boldsymbol{\beta}_{k+1}\right\| &\leq \left\|\int_0^1 \left(I_p - \delta\Lambda\left(\boldsymbol{\beta}^{\star} + t\Delta\boldsymbol{\beta}_k\right)\right)dt\Delta\boldsymbol{\beta}_k\right\| + \delta\left\|\eta_{1,n}\left(\boldsymbol{\beta}_k\right)\right\| + \delta\left\|\eta_{2,n}\right\| \\ &\leq \left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}\left(I_p - \delta\Lambda\left(\boldsymbol{\beta}\right)\right)\right\}\left\|\Delta\boldsymbol{\beta}_k\right\| + \delta\left\|\eta_{1,n}\left(\boldsymbol{\beta}_k\right)\right\| + \delta\left\|\eta_{2,n}\right\|.\end{aligned}$$

Note that for any $1 \leq s,t \leq p$,

$$\begin{aligned}\left|\left(\Lambda\left(\boldsymbol{\beta}\right)\right)_{s,t}\right| &= \left|\mathbb{E}\left[\int_{\mathcal{X}} \left(X_{s,i}X_{t,i} - X_{s,i}X_t\right)W\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)d\mathbf{X}\right]\right| \\ &\leq 2\left\|G'\right\|_{\infty}\mathbb{E}\left[\int_{\mathcal{X}} f_{\mathbf{X}|z}\left(\mathbf{X}|\,z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)d\mathbf{X}\right] = 2\left\|G'\right\|_{\infty},\end{aligned}$$

so each element of $\Lambda^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\Lambda\left(\boldsymbol{\beta}\right)$ is bounded by $2p\|G'\|_{\infty}$, and we have that

$$\begin{aligned}&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|\overline{\sigma}^2\left(I_p - \delta\Lambda\left(\boldsymbol{\beta}\right)\right) - \overline{\lambda}\left(I_p - \delta\left(\Lambda\left(\boldsymbol{\beta}\right) + \Lambda^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\right)\right)\right| \\ &\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\delta^2\left\|\Lambda^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\Lambda\left(\boldsymbol{\beta}\right)\right\| \leq 2\left\|G'\right\|_{\infty}p^2\delta^2.\end{aligned}$$

Then according to Assumption 2.5, we have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\sigma}^2\left(I_p - \delta\Lambda\left(\boldsymbol{\beta}\right)\right) \leq 1 - \delta\underline{\lambda}_{\Lambda} + 2\left\|G'\right\|_{\infty}p^2\delta^2.$$

When $\delta < \min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right), 1/\left(4\left\|G'\right\|_{\infty} p^2\right)\right\}$, we have that

$$0 \leq 1 - \delta\underline{\lambda}_{\Lambda} + 2\left\|G'\right\|_{\infty} p^2\delta^2 \leq 1 - \delta\underline{\lambda}_{\Lambda}/2 < 1.$$

So

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}\left(I_p - \delta\Lambda\left(\boldsymbol{\beta}\right)\right) \leq \sqrt{1 - \delta\underline{\lambda}_{\Lambda}/2} \leq 1 - \delta\underline{\lambda}_{\Lambda}/4,$$

and

$$\begin{aligned}
\left\|\Delta\boldsymbol{\beta}_{k+1}\right\| &\leq \left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)\left\|\Delta\boldsymbol{\beta}_k\right\| + \delta\left\|\eta_{1,n}\left(\boldsymbol{\beta}_k\right)\right\| + \delta\left\|\eta_{2,n}\right\| \\
&\leq \cdots \leq \left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| + \delta\cdot\sum_{j=0}^{k-1}\left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^j\left(\left\|\eta_{1,n}\left(\boldsymbol{\beta}_j\right)\right\| + \left\|\eta_{2,n}\right\|\right) \\
&\leq \left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| + \delta\cdot\sum_{j=0}^{\infty}\left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^j\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}\left(\boldsymbol{\beta}\right)\right\| + \left\|\eta_{2,n}\right\|\right) \\
&= \left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| + 4\underline{\lambda}_{\Lambda}^{-1}\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}\left(\boldsymbol{\beta}\right)\right\| + \left\|\eta_{2,n}\right\|\right).
\end{aligned}$$

Note that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}\left(\boldsymbol{\beta}\right)\right\| = p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n, p, h_n\right)$$

according to Lemma 1.1, and

$$\left\|\eta_{2,n}\right\| = O_p\left(\sqrt{p\left(\log p\right)/n}\right) = o_p\left(p^{\frac{5p+1}{2(p+1)}}\left(\psi\left(n, p, h_n\right)\right)^{\frac{1}{3p+3}}\right)$$

under any choices of $h_n \to 0$. This implies that when

$$\left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| \leq p^{\frac{5p+1}{2(p+1)}}\left(\psi\left(n, p, h_n\right)\right)^{\frac{1}{p+1}},$$

or equivalently,

$$k \geq k_{1,n}^{KBGD} = \frac{\log\left(\left\|\Delta\boldsymbol{\beta}_1\right\|\right) - \frac{5p+1}{2(p+1)}\log p - \frac{1}{p+1}\log\psi\left(n, p, h_n\right)}{-\log\left(1 - \delta\underline{\lambda}_{\Lambda}/4\right)},$$

we have that $\sup_{k\geq k_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_k\right\| = O_p\left(p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n, p, h_n\right)\right)$. □

Proof of Theorem 1.4

Proof. We first note that

$$\left\| \int_{\mathcal{X}} V\left(\mathbf{X}_{e,i}, \mathbf{X}_e, \boldsymbol{\beta}\right) d\mathbf{X} \right\| \leq 2p \left\| G' \right\|_{\infty} \int_{\mathcal{X}} f_{\mathbf{X}|z}\left(\mathbf{X} | \, z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right) d\mathbf{X} = 2p \left\| G' \right\|_{\infty},$$

for all $\mathbf{X}_{e,i}$, so

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \Lambda_{\phi}\left(\boldsymbol{\beta}\right) - \Lambda\left(\boldsymbol{\beta}\right) \right\| \leq 2p \left\| G' \right\|_{\infty} \mathbb{E}\left(1 - I_i^{\phi}\right) \leq 2\zeta p^2 \left\| G' \right\|_{\infty} \phi,$$

where the last inequality comes from the fact that $m\left(\mathcal{X}_e^{\phi}\right) = 1 - (1-\phi)^p \leq p\phi$. So

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \Lambda_{\phi}\left(\boldsymbol{\beta}\right) - \Lambda\left(\boldsymbol{\beta}\right) \right\| \leq \delta \underline{\lambda}_{\Lambda} / 8 \tag{1.12}$$

holds under the choice of ϕ.

Based on (1.12), the following proof is similar to the proof of Theorem 1.3. Define

$$\eta_{1,n}^{\phi}\left(\boldsymbol{\beta}\right) = \frac{1}{n} \sum_{i=1}^{n} \widehat{G}\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right) | \, \boldsymbol{\beta}\right) \mathbf{X}_i^{\phi} - \mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}\right), \boldsymbol{\beta}\right) \mathbf{X}_i^{\phi} \right],$$

$$\eta_{2,n}^{\phi} = \frac{1}{n} \sum_{i=1}^{n} G\left(z_i^{\star}\right) \mathbf{X}_i^{\phi} - \mathbb{E}\left[G\left(z_i^{\star}\right) \mathbf{X}_i^{\phi} \right] + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mathbf{X}_i^{\phi}.$$

We have that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+1} &= \Delta\boldsymbol{\beta}_k - \frac{\delta}{n} \sum_{i=1}^{n} \left(\widehat{G}\left(z_{i,k} | \, \boldsymbol{\beta}_k\right) - Y_i \right) \mathbf{X}_i^{\phi} \\
&= \Delta\boldsymbol{\beta}_k - \delta \mathbb{E}\left[\left(L\left(z_{i,k}, \boldsymbol{\beta}_k\right) - G\left(Z_i^{\star}\right) \right) \mathbf{X}_i^{\phi} \right] + \delta \left(\eta_{1,n}^{\phi}\left(\boldsymbol{\beta}_k\right) + \eta_{2,n}^{\phi} \right) \\
&= \int_0^1 \left\{ I_p - \delta \Lambda_{\phi}\left(\boldsymbol{\beta}^{\star} + t\Delta\boldsymbol{\beta}_k\right) \right\} \Delta\boldsymbol{\beta}_k dt + \delta \left(\eta_{1,n}^{\phi}\left(\boldsymbol{\beta}_k\right) + \eta_{2,n}^{\phi} \right),
\end{aligned}$$

so

$$\left\| \boldsymbol{\beta}_{k+1} \right\| \leq \sup_{\boldsymbol{\beta} \in \mathcal{B}} \overline{\sigma}\left(I_p - \delta \Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right) \left\| \boldsymbol{\beta}_k \right\| + \delta \left(\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \eta_{1,n}^{\phi}\left(\boldsymbol{\beta}\right) \right\| + \left\| \eta_{2,n}^{\phi} \right\| \right).$$

Obviously, since p is fixed, we have that $\left\| \eta_{2,n}^{\phi} \right\| = O_p\left(n^{-1/2}\right)$. Due to trimming, we also have that

$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}^{\phi}\left(\boldsymbol{\beta}\right)\right\|=O_{p}\left(\psi\left(n,p,h_{n}\right)\right)$. Note that (1.12) holds, so we have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\left\{I_{p}-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right\}-\left\{I_{p}-\delta\Lambda\left(\boldsymbol{\beta}\right)\right\}\right\|\leq\delta\underline{\lambda}_{\Lambda}/8.$$

According to the proof of Theorem 1.3, there holds $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\sigma}\left(I_{p}-\delta\Lambda\left(\boldsymbol{\beta}\right)\right)\leq1-\delta\underline{\lambda}_{\Lambda}/4$ under the choice of δ, so we have that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\sigma}\left(I_{p}-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}\right)\right)\leq1-\delta\underline{\lambda}_{\Lambda}/8.$$

Then based on the proof of Theorem 1.3, it remains to note that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left(\left\|\eta_{1,n}^{\phi}\left(\boldsymbol{\beta}\right)\right\|+\left\|\eta_{2,n}^{\phi}\right\|\right)=O_{p}\left(\psi\left(n,p,h_{n}\right)\right)$$

holds under any fixed trimming parameter ϕ. □

Proof of Lemma 2.3

Proof. Note that under the choice of δ and ϕ , $\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1}\left\|\boldsymbol{\beta}_{k}-\boldsymbol{\beta}^{\star}\right\|=O_{p}\left(\psi\left(n,p,h_{n}\right)\right)$ according to Theorem 1.4. According to (1.14), we have that

$$\begin{aligned}&\left\|\Delta\boldsymbol{\beta}_{k+\widetilde{k}_{1,n}^{KBGD}+1}\right\|\\&\leq\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1,,t\in[0,1]}\overline{\sigma}\left\{I_{p}-\frac{\delta}{n}\sum_{i=1}^{n}\left[\mathbf{X}_{i}^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_{k}}\right]\right\}\left\|\Delta\boldsymbol{\beta}_{k}\right\|+\delta\left\|\boldsymbol{\xi}_{n}^{\phi}\right\|.\end{aligned}$$

According to Lemma 1.2, we have that

$$\begin{aligned}&\sup_{k\geq\widetilde{k}_{1,n}^{KBGD},t\in[0,1]}\left\|\left\{I_{p}-\frac{\delta}{n}\sum_{i=1}^{n}\left[\mathbf{X}_{i}^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_{k}}\right]\right\}\right.\\&\left.-\left\{I_{p}-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_{k}\right)\right\}\right\|=\delta O_{p}\left(h_{n}^{-2}\sqrt{\left(\log\left(nh_{n}^{-1}\right)\right)/n}+h_{n}^{3}\right),\end{aligned}\tag{1.13}$$

due to the fact that

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_{k}\right\|=O_{p}\left(\psi_{1}\left(n,p,h_{n}\right)\right)=o_{p}\left(h_{n}^{-2}\sqrt{\left(\log\left(nh_{n}^{-1}\right)\right)/n}+h_{n}^{3}\right),$$

when p is fixed and $h_{n}\to0$.

When $nh_n^6 \to 0$ and $h_n^4 n/\left(\log n\right)^2 \to \infty$, we have that $h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n} + h_n^3 \to 0$. So we have that (1.13) is smaller than $\delta\underline{\lambda}_{\Lambda}/16$ with probability going to 1. According to the choice of ϕ and δ, we have that $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}\left(I_p - \delta\Lambda_\phi\left(\boldsymbol{\beta}\right)\right) \leq 1 - \delta\underline{\lambda}_{\Lambda}/8$ according to the proof of Theorem 1.4. So as n increases, with probability going to 1, there holds

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1, t\in[0,1]} \overline{\sigma}\left(I_p - \frac{\delta}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right]\right) \leq 1 - \delta\underline{\lambda}_{\Lambda}/16,$$

Then as n increases, with probability going to 1 there holds

$$\begin{aligned}\left\|\Delta\boldsymbol{\beta}_{k+\widetilde{k}_{1,n}^{KBGD}+1}\right\| &\leq \left(1-\delta\underline{\lambda}_{\Lambda}/16\right)\left\|\Delta\boldsymbol{\beta}_{k+\widetilde{k}_{1,n}^{KBGD}}\right\| + \delta\left\|\boldsymbol{\xi}_n^{\phi}\right\| \\ &\leq \cdots \leq \left(1-\delta\underline{\lambda}_{\Lambda}/16\right)^k\left\|\Delta\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD}+1}\right\| + 16\underline{\lambda}_{\Lambda}^{-1}\left\|\boldsymbol{\xi}_n^{\phi}\right\|.\end{aligned}$$

According to Lemma 1.3, $\left\|\boldsymbol{\xi}_n^{\phi}\right\| = O_p\left(n^{-1/2}\right)$. Also note that $\left\|\Delta\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD}+1}\right\| = O_p\left(\psi\left(n,p,h_n\right)\right)$, then if we choose $k_{2,n}^{KBGD}$ such that $\left(1-\delta\underline{\lambda}_{\Lambda}/16\right)^{k_{2,n}^{KBGD}-1} \leq n^{-1/2}\psi^{-1}\left(n,p,h_n\right)$, or equivalently,

$$k_{2,n}^{KBGD} \geq -\frac{\log\left(n^{1/2}\right)+\log\left(\psi\left(n,p,h_n\right)\right)}{\log\left(1-\delta\underline{\lambda}_{\Lambda}/16\right)} + 1,$$

we have that $\sup_{k\geq k_{2,n}^{KBGD}+1}\left\|\Delta\boldsymbol{\beta}_{k+\widetilde{k}_{1,n}^{KBGD}}\right\| = O_p\left(n^{-1/2}\right)$. This proves (i).

To prove (ii), we consider the following decomposition,

$$\Delta\boldsymbol{\beta}_{k+1} = \left(I_p - \delta\Lambda_\phi\left(\boldsymbol{\beta}^{\star}\right)\right)\Delta\boldsymbol{\beta}_k + \delta\overline{\omega}_1\left(\boldsymbol{\beta}_k\right) + \delta\overline{\omega}_2\left(\boldsymbol{\beta}_k\right) - \delta\boldsymbol{\xi}_n^{\phi},$$

where

$$\overline{\omega}_1\left(\boldsymbol{\beta}_k\right) = \int_0^1\left\{\Lambda_\phi\left(\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k\right) - \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^{\phi}\,\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right]\right\}dt\Delta\boldsymbol{\beta}_k,$$

and

$$\overline{\omega}_2\left(\boldsymbol{\beta}_k\right) = \int_0^1\left\{\Lambda_\phi\left(\boldsymbol{\beta}^{\star}\right) - \Lambda_\phi\left(\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k\right)\right\}dt\Delta\boldsymbol{\beta}_k.$$

Obviously, according to Lemma 1.2,

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+k_{2,n}^{KBGD}+1}\left\|\overline{\omega}_1\left(\boldsymbol{\beta}_k\right)\right\| = O_p\left(h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n}+h_n^3\right)O_p\left(n^{-\frac{1}{2}}\right) = o_p\left(n^{-\frac{1}{2}}\right).$$

We also note that each element of matrix $I_i^{\phi} \cdot \int_{\mathcal{X}} V(\mathbf{X}_{e,i}, \mathbf{X}_e, \boldsymbol{\beta})\, d\mathbf{X}$ has bounded derivative with respect to $\boldsymbol{\beta}$ for any $\mathbf{X}_{e,i}$. This is because, if $\mathbf{X}_{e,i} \notin \mathcal{X}_e^{\phi}$, $I_i^{\phi} = 0$ so each element will be zero and the results hold; if $\mathbf{X}_{e,i} \in \mathcal{X}_e^{\phi}$, then $f_z(z(\mathbf{X}_{e,i}, \boldsymbol{\beta}) | \boldsymbol{\beta}) > 0$, so $\int_{\mathcal{X}} \|\partial W(\mathbf{X}_{e,i}, \mathbf{X}_e, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}\|\, d\mathbf{X}$ is bounded according to Lemma 1.8(x). This implies that

$$\sup_{k \geq \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \|\overline{\omega}_2(\boldsymbol{\beta}_k)\| \leq C \|\Delta \boldsymbol{\beta}_k\|^2 = o_p\left(n^{-\frac{1}{2}}\right).$$

Then

$$\begin{aligned}
&\Delta \boldsymbol{\beta}_{k + \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \\
&= \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^k \Delta \boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} + \delta \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \overline{\omega}_1\left(\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + j}\right) \\
&+ \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \overline{\omega}_2\left(\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + j}\right) - \delta \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \boldsymbol{\xi}_n^{\phi}.
\end{aligned}$$

Note that $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \overline{\sigma}\left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta})\right) \leq 1 - \delta \underline{\lambda}_{\Lambda} / 8$, so

$$\begin{aligned}
\left\| \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^k \Delta \boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \right\| &\leq (1 - \delta \underline{\lambda}_{\Lambda} / 8)^k \left\| \Delta \boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \right\|, \\
\delta \left\| \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \omega_1\left(\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + j}\right) \right\| &\leq \delta \sum_{j=0}^{\infty} (1 - \delta \underline{\lambda}_{\Lambda} / 8)^j \sup_{k \geq \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + 1} \|\overline{\omega}_1(\boldsymbol{\beta}_k)\| \\
&= o_p\left(n^{-1/2}\right), \\
\delta \left\| \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \omega_2\left(\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD} + j}\right) \right\| &\leq \delta \sum_{j=0}^{\infty} (1 - \delta \underline{\lambda}_{\Lambda} / 8)^j \sup_{k \geq \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD}} \|\overline{\omega}_2(\boldsymbol{\beta}_k)\| \\
&= o_p\left(n^{-1/2}\right), \\
\left\| \Lambda_{\phi}^{-1}(\boldsymbol{\beta}^{\star}) \boldsymbol{\xi}_n^{\phi} - \delta \sum_{j=1}^{k} \left(I_p - \delta \Lambda_{\phi}(\boldsymbol{\beta}^{\star})\right)^{k-j} \boldsymbol{\xi}_n^{\phi} \right\| &\leq 8 \lambda_{\Lambda}^{-1} (1 - \delta \underline{\lambda}_{\Lambda} / 8)^{k+1} \left\| \boldsymbol{\xi}_n^{\phi} \right\|.
\end{aligned}$$

As $k \to \infty$, we have that $\lambda_{\Lambda}^{-1} (1 - \delta \underline{\lambda}_{\Lambda} / 8)^{k+1} \left\| \boldsymbol{\xi}_n^{\phi} \right\| = o_p\left(n^{-1/2}\right)$, so

$$\Delta \boldsymbol{\beta}_{k + \widetilde{k}_{1,n}^{KBGD} + k_{2,n}^{KBGD}} = \Lambda_{\phi}^{-1}(\boldsymbol{\beta}^{\star}) \boldsymbol{\xi}_n^{\phi} + o_p\left(n^{-1/2}\right).$$

According to Lemma 1.3, we have that $\sqrt{n}\boldsymbol{\xi}_n^{\phi} \to N\left(0, \Sigma_{\boldsymbol{\xi}}^{\phi}\right)$, so we have that

$$\sqrt{n}\Delta\boldsymbol{\beta}_{k+\widetilde{k}_{1,n}^{KBGD}+k_{2,n}^{KBGD}} = \Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\sqrt{n}\boldsymbol{\xi}_n^{\phi} + o_p\left(1\right) \to_d N\left(0, \Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\Sigma_{\boldsymbol{\xi}}^{\phi}\left(\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}\right).$$

□

Proof of Theorem 1.6

Proof. We only need to show that $\left\|\widehat{\Lambda}_{\phi,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right) - \Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| \to_p 0$ and $\left\|\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi} - \Sigma_{\boldsymbol{\xi}}^{\phi}\right\| \to_p 0$ both hold. Note that Lemma 1.2 indicates that $\left\|\widehat{\Lambda}_{\phi,n}\left(\widehat{\boldsymbol{\beta}}\right) - \Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right\| \to_p 0$, which implies that $\left\|\widehat{\Lambda}_{\phi,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right) - \Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right\| \to_p 0$ also holds.

Now we show that $\left\|\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi} - \Sigma_{\boldsymbol{\xi}}^{\phi}\right\| \to_p 0$ holds. Our basic proof method is similar to that of Lemma 1.1. In particular, let $\phi_n \downarrow 0$ and $\mathcal{X}_{e,n}$ be as defined as in the proof of Lemma 1.1. Then we have that $f_z^{\star}\left(z_i^{\star}\right) \geq C\phi_n^p$ as long as $\mathbf{X}_{e,i} \in \mathcal{X}_{e,n}$. Denote $G_i^{\star} = G\left(z_i^{\star}\right)$, we have

$$\begin{aligned}
\left\|\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi} - \Sigma_{\boldsymbol{\xi}}^{\phi}\right\| \leq & \left\|\frac{1}{n}\sum_{i=1}^{n}\left(I_{n,i}\cdot\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)^{\mathrm{T}}\right)\right. \\
& \left. -\mathbb{E}\left(I_{n,i}\cdot G_i^{\star}\left(1-G_i^{\star}\right)\left(\mathbf{X}_i^{\phi} - \mathbb{E}\left(\mathbf{X}_i^{\phi}\middle|z_i^{\star}\right)\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|z_i^{\star}\right)\right)^{\mathrm{T}}\right)\right\| \qquad (1.14) \\
& + \left\|\frac{1}{n}\sum_{i=1}^{n}\left((1-I_{n,i})\cdot\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)^{\mathrm{T}}\right)\right. \\
& \left. -\mathbb{E}\left((1-I_{n,i})\cdot G_i^{\star}\left(1-G_i^{\star}\right)\left(\mathbf{X}_i^{\phi} - \mathbb{E}\left(\mathbf{X}_i^{\phi}\middle|z_i^{\star}\right)\right)\left(\mathbf{X}_i^{\phi} - \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|z_i^{\star}\right)\right)^{\mathrm{T}}\right)\right\|. \qquad (1.15)
\end{aligned}$$

Note that $\widehat{G}_i$, $G_i^{\star}$, $\mathbf{X}_i^{\phi}$, $\widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)$, and $\mathbb{E}\left(\mathbf{X}_i^{\phi}\middle|z_i^{\star}\right)$ are all upper bounded, so (1.15) is $O_p\left(\phi_n\right)$. Now we look at (1.14). Note that

$$\widehat{G}_i - \frac{\sum_{j=1}^{n}K_{h_n}\left(z_i^{\star}-z_j^{\star}\right)y_j}{\sum_{j=1}^{n}K_{h_n}\left(z_i^{\star}-z_j^{\star}\right)} = \frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\widetilde{\boldsymbol{\beta}}\right)\middle|\widetilde{\boldsymbol{\beta}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\Delta\widehat{\boldsymbol{\beta}},$$

where $\widetilde{\boldsymbol{\beta}}$ lies somewhere between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\star}$. According to the proof of Lemma 3.1, we have that

$$\sup_{\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\in\mathcal{X}_{e,n}\times\mathcal{B}}\left\|\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right\| = O_p\left(1\right)$$

if $\phi_n^{-p}\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right) \to 0$, since $\left\|f_z^{-1}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\partial H_1\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z\right\|$ and $\left\|L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)f_z^{-1}\left(z\left(\mathbf{X}_e,\right.\right.\right.$

are both bounded for all $\boldsymbol{\beta} \in \mathcal{B}$ and $\mathbf{X}_e \in \mathcal{X}_{e,n}$. So

$$\max_{1\leq i\leq n}\left|\left(\widehat{G}_i - \frac{\sum_{j=1}^n K_{h_n}\left(z_i^\star - z_j^\star\right) y_j}{\sum_{j=1}^n K_{h_n}\left(z_i^\star - z_j^\star\right)}\right)\cdot I_{n,i}\right| = O_p\left(n^{-1/2}\right).$$

Also note that when $\phi_n^{-p}\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right) \to 0$,

$$\max_{1\leq i\leq n}\left|\left(\frac{\sum_{j=1}^n K_{h_n}\left(z_i^\star - z_j^\star\right) y_j}{\sum_{j=1}^n K_{h_n}\left(z_i^\star - z_j^\star\right)} - G\left(z_i^\star\right)\right)\cdot I_{n,i}\right| = O_p\left(\phi_n^{-p}\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right)\right),$$

this indicates that

$$\max_{1\leq i\leq n} I_{n,i}\cdot\left|\widehat{G}_i - G\left(z_i^\star\right)\right| = O_p\left(\phi_n^{-p}\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right)\right),$$

due to $n^{1/2}\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right) \to \infty$ under the choice of h_n. Using similar argument, we can also show that

$$\max_{1\leq i\leq n}\left\|\left(\widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi\middle|\widehat{z}_i\right) - \mathbb{E}\left(\mathbf{X}_i^\phi\middle|z_i^\star\right)\right)\cdot I_{n,i}\right\| = O_p\left(\phi_n^{-p}\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right)\right).$$

So we have that (1.14) is of order $O_p\left(\phi_n^{-p}\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right) + n^{-1/2}\right)$. It remains to choose

$$\phi_n = O\left(\left(h_n^{-1}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right)^{\frac{1}{p+1}}\right)$$

to conclude the proof. □

Proof of Theorem 1.7

Proof. The proof is similar to that of Theorem 1.3. Note that

$$\begin{aligned}&\sup_{0\leq t\leq 1,\boldsymbol{\beta}\in\mathcal{B}}\left|\overline{\sigma}^2\left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}\right)\right) - \overline{\lambda}\left(I_p - \delta\left(\Psi_q\left(t,\boldsymbol{\beta}\right) + \Psi_q^{\mathrm{T}}\left(t,\boldsymbol{\beta}\right)\right)\right)\right|\\ &\leq \delta^2\sup_{0\leq t\leq 1,\boldsymbol{\beta}\in\mathcal{B}}\left\|\Psi_q\left(t,\boldsymbol{\beta}\right)\right\|^2 \leq \delta^2\left\|G'\right\|_\infty^2 p^2\left\{1 + \underline{\lambda}_\Gamma^{-1} q D_{q,0}^2\right\}^2.\end{aligned}$$

So if $\delta^2 \left\|G'\right\|_\infty^2 p^2 \left\{1+\underline{\lambda}_\Gamma^{-1} q D_{q,0}^2\right\}^2 \leq \frac{1}{2}\underline{\lambda}_\Psi \delta$, or equivalently, $\delta \leq \underline{\lambda}_\Psi / \left(2\left\|G'\right\|_\infty^2 p^2 \left\{1+\underline{\lambda}_\Gamma^{-1} q D_{q,0}^2\right\}^2\right)$, we have that

$$\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}} \left|\overline{\sigma}^2\left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}\right)\right) - \overline{\lambda}\left(I_p - \delta\left(\Psi_q\left(t,\boldsymbol{\beta}\right) + \Psi_q^{\mathrm{T}}\left(t,\boldsymbol{\beta}\right)\right)\right)\right| \leq \underline{\lambda}_\Psi \delta/2,$$

so

$$\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}^2\left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}\right)\right) \leq 1 - \underline{\lambda}_\Psi \delta/2 < 1,$$

and

$$\sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}\left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}\right)\right) \leq 1 - \underline{\lambda}_\Psi \delta/4.$$

Then we have that

$$\begin{aligned}
\left\|\Delta\boldsymbol{\beta}_{k+1}\right\| &\leq \left\|\int_0^1 \left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}_k\right)\right)\Delta\boldsymbol{\beta} dt + \delta\mathfrak{R}_{n,k}\right\| \\
&\leq \sup_{0\leq t\leq 1, \boldsymbol{\beta}\in\mathcal{B}} \overline{\sigma}\left(I_p - \delta\Psi_q\left(t,\boldsymbol{\beta}\right)\right)\left\|\Delta\boldsymbol{\beta}_k\right\| + \delta_k\left\|\mathfrak{R}_{n,k}\right\| \leq \left(1-\underline{\lambda}_\Psi\delta/4\right)\left\|\Delta\boldsymbol{\beta}_k\right\| + \delta\left\|\mathfrak{R}_{n,k}\right\| \leq \cdots \\
&\leq \left(1-\underline{\lambda}_\Psi\delta/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| + \delta\sum_{j=1}^k \left(1-\underline{\lambda}_\Psi\delta/4\right)^{k-j}\left\|\mathfrak{R}_{n,j}\right\| \\
&\leq \left(1-\underline{\lambda}_\Psi\delta/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| + 4/\underline{\lambda}_\Psi O_p\left(\sup_{k\geq 1}\left\|\mathfrak{R}_{n,k}\right\|\right).
\end{aligned}$$

When $\left(1-\underline{\lambda}_\Psi\delta/4\right)^k\left\|\Delta\boldsymbol{\beta}_1\right\| \leq \chi_{2,n}$, or equivalently, $k \geq \frac{\log(\|\Delta\boldsymbol{\beta}_1\|)-\log(\chi_{2,n})}{-\log\left(1-\underline{\lambda}_\Psi\delta/4\right)} = k_{1,n}^{SBGD}$, there holds $\left\|\Delta\boldsymbol{\beta}_{k+1}\right\| = O_p\left(\chi_{2,n}\right)$. □

Proof of Theorem 1.8

Proof. We first prove Theorem 1.8 (i). Note that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+1} &= \left\{\int_0^1 \left(I_p - \delta\Psi_q^\star\right)dt\right\}\Delta\boldsymbol{\beta}_k + \delta\mathfrak{R}_{n,k} \\
&= \left(I_p - \delta\Psi_q^\star\right)\Delta\boldsymbol{\beta}_k + \frac{\delta}{n}\sum_{i=1}^n \left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i + \delta\left\{\mathfrak{R}_{n,k} - \frac{1}{n}\sum_{i=1}^n \left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i\right. \\
&\left. + \int_0^1 \left(\Psi_q^\star - \frac{1}{n}\sum_{i=1}^n G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)\right)dt\Delta\boldsymbol{\beta}_k\right\}.
\end{aligned}$$

Define

$$\widetilde{\mathfrak{R}}_{n,k} = \mathfrak{R}_{n,k} - \frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i + \\ \int_0^1 \left(\Psi_q^\star - \frac{1}{n}\sum_{i=1}^{n} G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)\right) dt\Delta\boldsymbol{\beta}_k.$$

According to Lemma 1.5, we have that

$$\begin{aligned}
&\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\int_0^1 \left(\Psi_q^\star - \frac{1}{n}\sum_{i=1}^{n} G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)\right) dt\Delta\boldsymbol{\beta}_k\right\| \\
&\leq \sup_{k\geq k_{1,n}^{SBGD}+1,0\leq t\leq 1}\left\|\Psi_q^\star - \frac{1}{n}\sum_{i=1}^{n} G'\left(z_i^\star + t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)\right\| \sup_{k\geq k_{1,n}^{SBGD}+1}\|\Delta\boldsymbol{\beta}_k\| \\
&= O_p\left(\sqrt{p}qD_{q,0}^2\left(p + qD_{q,0}D_{q,1}\right)\sup_{k\geq k_{1,n}^{SBGD}+1}\|\Delta\boldsymbol{\beta}\|^2\right) \\
&= O_p\left(\sqrt{p}qD_{q,0}^2\left(p + qD_{q,0}D_{q,1}\right)\chi_{2,n}^2\right).
\end{aligned}$$

According to Lemma 1.6, we have that

$$\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\mathfrak{R}_{n,k} - \frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i\right\| = O_p\left(\chi_{4,n}\right).$$

This shows the result.

To prove Theorem 1.8(ii), we note that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+k_{1,n}^{SBGD}+1} &= \left(I_p - \delta\Psi_q^\star\right)\Delta\boldsymbol{\beta}_{k+k_{1,n}^{SBGD}} + \frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i + \widetilde{\mathfrak{R}}_{n,k+k_{1,n}^{SBGD}}, \\
&= \left(I_p - \delta\Psi_q^\star\right)^k\Delta\boldsymbol{\beta}_{k_{1,n}^{SBGD}+1} + \sum_{j=1}^{k}\left(I_p - \delta\Psi_q^\star\right)^{j-1}\left(\frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i\right) \\
&+ \sum_{j=1}^{k}\left(I_p - \delta\Psi_q^\star\right)^{j-1}\widetilde{\mathfrak{R}}_{n,k+k_{1,n}^{SBGD}+1-j} \\
&= \Psi_q^{\star-1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i + \left(I_p - \delta\Psi_q^\star\right)^k\Delta\boldsymbol{\beta}_{k_{1,n}^{SBGD}+1} + \sum_{j=1}^{k}\left(I_p - \delta\Psi_q^\star\right)^{j-1}\widetilde{\mathfrak{R}}_{n,k+k_{1,n}^{SBGD}+1-j} \\
&+ \sum_{j=k+1}^{\infty}\left(I_p - \delta\Psi_q^\star\right)^{j-1}\left(\frac{\delta}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i - \mathfrak{X}_{q,i}\right)\varepsilon_i\right).
\end{aligned}$$

Then since

$$\left\|\left(I_p - \delta\Psi_q^\star\right)^k\Delta\boldsymbol{\beta}_{k_{1,n}^{SBGD}+1}\right\| = O_p\left(\left(1 - \underline{\lambda}_\Psi\delta/4\right)^k\chi_{2,n}\right),$$

$$\left\|\sum_{j=1}^{k}\left(I_p-\delta\Psi_q^{\star}\right)^{j-1}\widetilde{\mathfrak{R}}_{n,k+k_{1,n}^{SBGD}+1-j}\right\|\leq\sum_{j=1}^{\infty}\left(1-\underline{\lambda}_{\Psi}\delta/4\right)^{j-1}\sup_{k\geq k_{1,n}^{SBGD}+1}\left\|\widetilde{\mathfrak{R}}_{n,k}\right\|=O_p\left(\chi_{5,n}\right),$$

and

$$\begin{aligned}\left\|\sum_{j=k+1}^{\infty}\left(I_p-\delta\Psi_q^{\star}\right)^{j-1}\left(\frac{\delta}{n}\sum_{i=1}^{n}\left(\mathfrak{V}_q\boldsymbol{r}_q\left(z_i^{\star}\right)+\mathbf{X}_i\right)\varepsilon_i\right)\right\|&\leq\left(1-\underline{\lambda}_{\Psi}\delta/4\right)^{k}\left\|\frac{4}{\underline{\lambda}_{\Psi}n}\sum_{i=1}^{n}\left(\mathbf{X}_i-\mathfrak{X}_{q,i}\right)\varepsilon_i\right\|\\&=O_p\left(\left(1-\underline{\lambda}_{\Psi}\delta/4\right)^{k}\sqrt{\frac{pqD_{q,0}^2\left(\log p\right)}{n}}\right)\\&=O_p\left(\left(1-\underline{\lambda}_{\Psi}\delta/4\right)^{k}\chi_{2,n}\right).\end{aligned}$$

So as long as $\left(1-\underline{\lambda}_{\Psi}\delta/4\right)^{k}\chi_{2,n}\leq n^{-1/2}$, or equivalently, $k\geq k_{2,n}^{SBGD}=\frac{-\log\chi_{2,n}+\log\sqrt{n}}{-\log\left(1-\underline{\lambda}_{\Psi}\delta/4\right)}$, we have that

$$\sup_{k\geq k_{2,n}^{SBGD}+1}\left\|\Delta\boldsymbol{\beta}_{k+k_{1,n}^{SBGD}+1}-\Psi_q^{\star-1}\frac{1}{n}\sum_{i=1}^{n}\left(\mathfrak{V}_q\boldsymbol{r}_q\left(z_i^{\star}\right)+\mathbf{X}_i\right)\varepsilon_i\right\|=o_p\left(n^{-\frac{1}{2}}\right).$$

The following results hold trivially. □

Proof of Theorem 1.9

Proof. Note that under all the conditions imposed in Theorem 1.8, we have that

$$\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{\star}\right\|=O_p\left(\sqrt{pq^2D_{q,0}^4\left(\log p\right)/n}\right),$$

due to the fact that each element of $\left(\mathbf{X}_i-\mathfrak{X}_{q,i}\right)\varepsilon_i$ is bounded by $CqD_{q,0}^2$ and Assumption 1.7 holds.

To prove the theorem, we first show that

$$\sup_{1\leq i\leq n}\left|\widehat{G}_i-G_i\left(z_i^{\star}\right)\right|=O_p\left(\sqrt{p^2q^4D_{q,0}^8\left(\log p\right)/n}+qD_{q,0}^2\mathcal{E}_{q,0}\right).$$

Define $\widehat{z}_i=z\left(\mathbf{X}_{e,i},\widehat{\boldsymbol{\beta}}\right)$. To show the above result, note that

$$\begin{aligned}\sup_{1\leq i\leq n}\left|\widehat{G}_i-G\left(z_i^{\star}\right)\right|\leq&\sup_{1\leq i\leq n}\left|\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\left(\widehat{\boldsymbol{\pi}}_q-\boldsymbol{\pi}_q^{\star}\right)\right|\\&+\sup_{1\leq i\leq n}\left|\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\boldsymbol{\pi}_q^{\star}-G\left(\widehat{z}_i\right)\right|+\sup_{1\leq i\leq n}\left|G\left(\widehat{z}_i\right)-G\left(z_i^{\star}\right)\right|.\end{aligned}$$

Obviously, the second and third terms on RHS are of order $O_p\left(\mathcal{E}_{q,0}\right)$ and $O_p\left(\sqrt{p^2q^2D_{q,0}^4\left(\log p\right)/n}\right)$,

while the first term is bounded by $\sqrt{q}D_{q,0}\left\|\widehat{\boldsymbol{\pi}}_q-\boldsymbol{\pi}_q^\star\right\|$. Note that

$$\begin{aligned}\widehat{\boldsymbol{\pi}}_q-\boldsymbol{\pi}_q^\star=\Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\left(\frac{1}{n}\sum_{i=1}^n\widehat{\boldsymbol{r}}_{q,i}\left(G\left(\widehat{z}_i\right)-G\left(z_i^\star\right)\right)\right)+\Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\left(\frac{1}{n}\sum_{i=1}^n\widehat{\boldsymbol{r}}_{q,i}R_q\left(\widehat{z}_i\right)\right)\\+\Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{r}_q\left(\widehat{z}_i\right)\varepsilon_i\right).\end{aligned}$$

So we have that $\left\|\widehat{\boldsymbol{\pi}}_q-\boldsymbol{\pi}_q^\star\right\|=O_p\left(\sqrt{p^2q^3D_{q,0}^6\left(\log p\right)/n}+\sqrt{q}D_{q,0}\mathcal{E}_{q,0}\right)$ and the third term is of order $O_p\left(\sqrt{p^2q^4D_{q,0}^8\left(\log p\right)/n}+qD_{q,0}^2\mathcal{E}_{q,0}\right)$. This proves the first result.

We also note that according to the proof of Lemma 1.17, we have that

$$\sup_{1\leq i\leq n}\left\|\mathfrak{X}_{q,n}\left(\widehat{z}_i,\widehat{\boldsymbol{\beta}}\right)-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right\|=O_p\left(\sqrt{p^3q^6D_{q,0}^{10}D_{q,1}^2\log\left(pn\right)/n}\right).$$

Then we show that

$$\begin{aligned}&\max_{1\leq i\leq n}\left\|\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}-G_i\left(1-G_i\right)\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)^{\mathrm{T}}\right\|\\&=O_p\left(\sqrt{p^4q^8D_{q,0}^{14}\left(\log pn\right)/n}\left(D_{q,0}+D_{q,1}\right)+pq^3D_{q,0}^6\mathcal{E}_{q,0}\right).\end{aligned}$$

Note that the above is bounded by

$$\begin{aligned}&\max_{1\leq i\leq n}\left\|\left(\widehat{G}_i\left(1-\widehat{G}_i\right)-G_i\left(1-G_i\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right\|\\&+\max_{1\leq i\leq n}\left\|G_i\left(1-G_i\right)\left(\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}-\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)^{\mathrm{T}}\right)\right\|,\end{aligned}$$

where the first term is of order $O_p\left(\sqrt{p^4q^8D_{q,0}^{16}\left(\log p\right)/n}+pq^3D_{q,0}^6\mathcal{E}_{q,0}\right)$, while the second term is of order $O_p\left(\sqrt{p^4q^8D_{q,0}^{14}D_{q,1}^2\left(\log pn\right)/n}\right)$. Together we show the result.

Next we show that

$$\left\|\widehat{\Psi}_q^\star-\Psi_q^\star\right\|=O_p\left(\sqrt{p^4q^4D_{q,0}^4\log\left(pqD_{q,0}D_{q,1}n\right)/n}\right).$$

Since $\upsilon_G \geq 2$, we have that

$$\begin{aligned}
&\sup_{1\leq i\leq n}\left|\widehat{G}_i' - G'\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}^\star\right)\right)\right| \leq \sup_{1\leq i\leq n}\left|\widehat{\boldsymbol{r}}_{q,i}'^{\mathrm{T}}\left(\widehat{\boldsymbol{\pi}}_q - \boldsymbol{\pi}_q^\star\right)\right| \\
&+ \sup_{1\leq i\leq n}\left|\widehat{\boldsymbol{r}}_{q,i}'^{\mathrm{T}}\boldsymbol{\pi}_q^\star - G'\left(\widehat{z}_i\right)\right| + \sup_{1\leq i\leq n}\left|G'\left(\widehat{z}_i\right) - G'\left(z_i^\star\right)\right| \\
&= O_p\left(\sqrt{p^2q^4D_{q,0}^8D_{q,1}^2\left(\log p\right)/n} + qD_{q,0}D_{q,1}\mathcal{E}_{q,0} + \mathcal{E}_{q,1}\right).
\end{aligned}$$

So

$$\begin{aligned}
\left\|\widehat{\Psi}_q^\star - \Psi_q^\star\right\| \leq &\left\|\frac{1}{n}\sum_{i=1}^n\left(\widehat{G}_i' - G'\left(z_i^\star\right)\right)\cdot\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(\widehat{z}_i,\widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i^{\mathrm{T}}\right)\right\|, \\
&+\left\|\frac{1}{n}\sum_{i=1}^n G'\left(z_i^\star\right)\cdot\left(\left(\mathfrak{X}_{q,n}\left(\widehat{z}_i,\widehat{\boldsymbol{\beta}}\right) - \mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\mathbf{X}_i^{\mathrm{T}}\right)\right\| \\
&+\left\|\frac{1}{n}\sum_{i=1}^n G'\left(z_i^\star\right)\cdot\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\mathbf{X}_i^{\mathrm{T}} - \Psi_q^\star\right\| \\
&= O_p\left(\sqrt{p^4q^6D_{q,0}^{12}D_{q,1}^2\log\left(pn\right)/n} + pq^2D_{q,0}^3D_{q,1}\mathcal{E}_{q,0} + pqD_{q,0}^2\mathcal{E}_{q,1}\right),
\end{aligned}$$

which also implies that $\overline{\sigma}\left(\widehat{\Psi}_q^{\star-1}\right) = O_p\left(1\right)$, and

$$\left\|\widehat{\Psi}_q^{\star-1} - \Psi_q^{\star-1}\right\| = O_p\left(\sqrt{p^4q^6D_{q,0}^{12}D_{q,1}^2\left(\log pn\right)/n} + pq^2D_{q,0}^3D_{q,1}\mathcal{E}_{q,0} + qD_{q,0}^2\mathcal{E}_{q,1}\right)$$

Now we are ready to demonstrate the consistency of the variance estimator. Note that

$$\begin{aligned}
&\left|\widehat{\sigma}_S^2\left(\rho\right) - \sigma_S^2\left(\rho\right)\right| \\
&\leq \left\|\rho\right\|^2\left\|\widehat{\Psi}_q^{\star-1}\frac{1}{n}\sum_{i=1}^n\left\{\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right\}\left(\widehat{\Psi}_q^{\star-1}\right)^{\mathrm{T}}\right. \\
&\left. - \Psi_q^{\star-1}\mathbb{E}\left\{G\left(z_i^\star\right)\left(1-G\left(z_i^\star\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)^{\mathrm{T}}\right\}\left(\Psi_q^{\star-1}\right)^{\mathrm{T}}\right\| \\
&\leq \left\|\rho\right\|^2\left\|\widehat{\Psi}_q^{\star-1} - \Psi_q^{\star-1}\right\|\left\|\frac{1}{n}\sum_{i=1}^n\left\{\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right\}\left(\Psi_q^{\star-1}\right)^{\mathrm{T}}\right\| \\
&+ \left\|\rho\right\|^2\left\|\Psi_q^{\star-1}\left(\frac{1}{n}\sum_{i=1}^n\left\{\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right\}\right.\right. \\
&\left.\left. -\mathbb{E}\left\{G\left(z_i^\star\right)\left(1-G\left(z_i^\star\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_{q,n}\left(\widehat{z},\widehat{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}\right\}\right)\left(\widehat{\Psi}_q^{\star-1}\right)^{\mathrm{T}}\right\| \\
&+ \left\|\rho\right\|^2\left\|\Psi_q^{\star-1}\mathbb{E}\left\{G\left(z_i^\star\right)\left(1-G\left(z_i^\star\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}_i - \mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)^{\mathrm{T}}\right\}\left(\widehat{\Psi}_q^{\star-1} - \Psi_q^{\star-1}\right)\right\|.
\end{aligned}$$

The first and the third terms are of order $O_p\left(\sqrt{p^6q^8D_{q,0}^{16}D_{q,1}^2\left(\log pn\right)/n} + p^2q^3D_{q,0}^4D_{q,1}\mathcal{E}_{q,0} + pq^2D_{q,0}^4\mathcal{E}_{q,1}\right)$,

and the second term is of order $O_p\left(\sqrt{p^4q^8D_{q,0}^{14}\left(\log pn\right)/n}\left(D_{q,0}+D_{q,1}\right)+pq^3D_{q,0}^6\mathcal{E}_{q,0}\right)$. Together, we have that

$$\left|\widehat{\sigma}_S^2\left(\rho\right)-\sigma_S^2\left(\rho\right)\right|=O_p\left(\sqrt{p^6q^8D_{q,0}^{16}D_{q,1}^2\left(\log pn\right)/n}+pq^3D_{q,0}^4\left(pD_{q,1}+D_{q,0}^2\right)\mathcal{E}_{q,0}+pq^2D_{q,0}^4\mathcal{E}_{q,1}\right),$$

which implies that $\left|\widehat{\sigma}_S^2\left(\rho\right)-\sigma_S^2\left(\rho\right)\right|\to_p 0$ under all the conditions. □

Chapter 2

Stochastic Learning of Semiparametric Monotone Index Models with Large Sample Size

2.1 Introduction

Consider estimating the unknown parameter in a monotone index model (Han, 1987; Cavanagh and Sherman, 1998) using loss function $L(\mathbf{X}_e, y, \boldsymbol{\beta}_e|G)$ which is differentiable with respect to $\boldsymbol{\beta}_e$ and satisfies $\boldsymbol{\beta}_e^\star = \arg\min_{\boldsymbol{\beta}_e \in \mathcal{B}_e} \mathbb{E}[L(y, \mathbf{X}_e, \boldsymbol{\beta}_e|G)]$, where y is the response variable, $\mathbf{X}_e = \left(X_0, \mathbf{X}^{\mathrm{T}}\right)^{\mathrm{T}} = (X_0, X_1, \cdots, X_p)^{\mathrm{T}} \in \mathcal{X}_e$ is the covariate, $\boldsymbol{\beta}_e^\star = \left(\beta_0^\star, \boldsymbol{\beta}^{\star\mathrm{T}}\right)^{\mathrm{T}} = \left(\beta_0^\star, \beta_1^\star, \cdots, \beta_p^\star\right)^{\mathrm{T}} \in \mathcal{B}_e$ is the true parameter, and $G(z) = \mathbb{E}(y|\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e^\star = z)$ is the monotone link function. When G is known, given data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$, the estimator of $\boldsymbol{\beta}_e^\star$ can be constructed as $\widehat{\boldsymbol{\beta}}_{e,n} = \arg\min_{\boldsymbol{\beta}_e \in \mathcal{B}_e} \frac{1}{n}\sum_{i=1}^n L(\mathbf{X}_{e,i}, y_i, \boldsymbol{\beta}_e|G)$. However, when G is unknown, the estimation problem becomes semiparametric, and the above estimator can not be constructed directly. One solution is to replace the unknown G with its Nadaraya-Watson estimator

$$\widehat{G}_n(z|\boldsymbol{\beta}_e) = \frac{\frac{1}{nh_n}\sum_{j=1}^n K((z - \mathbf{X}_{e,j}^{\mathrm{T}}\boldsymbol{\beta}_e)/h_n) y_j}{\frac{1}{nh_n}\sum_{j=1}^n K((z - \mathbf{X}_{e,j}^{\mathrm{T}}\boldsymbol{\beta}_e)/h_n)},$$

where K is the kernel function and h_n is the bandwidth parameter. Then the estimator of $\boldsymbol{\beta}_e^\star$ can be constructed by minimizing the new plug-in loss function

$$\widehat{\boldsymbol{\beta}}_{e,n} = \arg\min_{\boldsymbol{\beta}_e \in \mathcal{B}_e} \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{X}_{e,i}, y_i, \boldsymbol{\beta}_e | \widehat{G}_n(\cdot|\boldsymbol{\beta}_e)). \tag{2.1}$$

Ichimura (1993); Härdle et al. (1993); Klein and Spady (1993); Rothe (2009) consider different loss funcctions and provide sufficient conditions that guarantee $1/\sqrt{n}$-consistency and asymptotic normality of $\widehat{\boldsymbol{\beta}}_{e,n}$ in (2.1).

This paper investigates the *computational* aspect of semiparametric estimation (2.1) under an extremely large n setup. In this scenario, hundreds of thousands of or even millions of data points are available for estimation. Indeed, due to the rapid development of technology in data collection and data storage, it's becoming more and more common nowadays for data analysts to deal with data set with extraordinary amount of observations (Wang et al., 2018). This offers the researchers opportunities to more precisely understand the potential mechanism lurking behind the data, while on the same time brings about a series of new challenges. Among others, the key challenge is the extremely heavy computational burdens and exhaustive computational time that make the existing statistical methods numerically prohibitive. Consequently, it's more urgent than ever before to study estimation methods that is applicable in the big-data era. In the recent literature, many methods for large n estimation have been extensively studied, including subsample-based optimization (Forneron, 2022; Toulis and Airoldi, 2017) and estimation (Wang et al., 2018; Wang, 2019).

To get a brief idea of the computational difficulty associated with the semiparametric estimator (2.1), we note that to numerically solve the optimization problem in (2.1), gradient-based methods are generally applied. In particular, starting from an initial guess of $\widehat{\boldsymbol{\beta}}_{e,n}$, $\boldsymbol{\beta}_{e,1}$, the following Batch Gradient Descent (BGD) iterations are repeatedly performed until some terminating conditions are satisfied (Ruder, 2016; Bottou et al., 2018)

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n}\sum_{i=1}^{n} \frac{\partial L(\mathbf{X}_{e,i}, y, \boldsymbol{\beta}_{e,k} | \widehat{G}_n(\cdot|\boldsymbol{\beta}_{e,k}))}{\partial \boldsymbol{\beta}_e}, \tag{2.2}$$

where $\delta_k > 0$ is the learning rate. Note that the gradient of $L(\mathbf{X}_e, y, \boldsymbol{\beta}_e | \widehat{G}_n(\cdot|\boldsymbol{\beta}_e))$ with respect to $\boldsymbol{\beta}_e$ generally depends on $\widehat{G}_n(\cdot|\boldsymbol{\beta}_e)$, $\partial \widehat{G}_n(z|\boldsymbol{\beta}_e)/\partial z$, and $\partial \widehat{G}_n(\cdot|\boldsymbol{\beta}_e)/\partial \boldsymbol{\beta}_e$. For example, let $L(\mathbf{X}_e, y, \boldsymbol{\beta}_e|G) =$

$(y - G(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e))^2$ as in Ichimura (1993), there obviously holds

$$\begin{aligned}&\frac{\partial L(\mathbf{X}_e, y, \boldsymbol{\beta}_e|\widehat{G}_n(\cdot|\boldsymbol{\beta}_e))}{\partial \boldsymbol{\beta}_e}\\&= 2(\widehat{G}_n(\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e|\boldsymbol{\beta}_e) - y_i)\left(\left.\frac{\partial \widehat{G}_n(z|\boldsymbol{\beta}_e)}{\partial z}\right|_{z=\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e}\mathbf{X}_e + \left.\frac{\partial \widehat{G}_n(z|\boldsymbol{\beta}_e)}{\partial \boldsymbol{\beta}_e}\right|_{z=\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e}\right).\end{aligned}$$

Due to the nature of Nadaraya-Watson kernel estimator, for each input z, evaluating $\widehat{G}_n(z|\boldsymbol{\beta}_e)$, $\partial\widehat{G}_n(z|\boldsymbol{\beta}_e)/\partial z$, and $\partial\widehat{G}_n(z|\boldsymbol{\beta}_e)/\partial\boldsymbol{\beta}_e$ involves performing summations over n data points, so requires computational time of order $O(n)$. This implies that evaluating the gradient of the plug-in loss function at merely one data point requires computational time of order $O(n)$. Consequently, evaluating the gradient functions at all data points and performing a single update based on (2.2) require computational time of order $O(n^2)$. Such computational complexity increases too fast with the sample size n, and makes, as was pointed out by Ichimura (1993), solving (2.1) roughly n times more numerically complicated compared with worst-case parametric estimation with differentiable loss function[1]. This renders semiparametric estimation numerically infeasible even in the modest n scenario.

Nevertheless, it is straightforward to use stochastic optimization strategies to alleviate the devastating computational burden of the above semiparametric estimation when n is extremely large (Ruder, 2016; Bottou et al., 2018). For example, instead of using the full data set to perform the update in (2.2), one may resort to using only a subset of the data. In particular, Mini-Batch Gradient Descent (MBGD) suggests the following update (Ruder, 2016; Bottou et al., 2018; Forneron, 2022),

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{B}\sum_{i\in\mathfrak{I}_{B,k}}\frac{\partial L(\mathbf{X}_{e,i}, y_i, \boldsymbol{\beta}_{e,k}|\widehat{G}_n(\cdot|\boldsymbol{\beta}_{e,k}))}{\partial \boldsymbol{\beta}_e}, \tag{2.3}$$

where B is the subsample size and $\mathfrak{I}_{B,k}$ is the subsample index set that is randomly drawn from $\{1, 2, \cdots, n\}$ with replacement and is independent over k. Under MBGD algorithm, the gradient of the plug-in loss function is evaluated over B data points, so the computational time for each update is of order $O(nB)$. Obviously, $nB \ll n^2$ if we choose $B/n \to 0$, so the computational burden is relieved to some extent if we choose B diverging more slowly than n. While when pursuing $1/\sqrt{n}$-consistent estimators, it's generally required that $B/\sqrt{n} \to \infty$ (Forneron, 2022), so the updating time is still of order at least $O(n\sqrt{n})$, which is $\sqrt{n}$ times more complicated than the worst-case

[1]When the loss function is smooth and differentiable with respect to the parameter, each single update of the parameter based on BGD algorithm requires computational time of order $O(n)$ (Ruder, 2016; Bottou et al., 2018).

parametric estimation with differentiable loss function. This is still numerically difficult when n is extremely large. Moreover, different from the conventional stochastic optimization approach that imposes less pressure on the computer memory requirement[2], in the above semiparametric setup, even if stochastic optimization (2.3) is applied, all the data points have to be stored in the memory because each evaluation of the kernel estimator (or its gradient with respect to $\boldsymbol{\beta}_e$) requires access to all the data points. This imposes heavy burden on the computer memory when there are millions or even trillions of data points.

Motivated by the extremely heavy computational burdens and intensive memory requirement caused by large sample size, this paper proposes a novel computationally friendly estimation procedure for semiparametric monotone index models. We propose a new iterative algorithm whose computational complexity for each update can be made sufficiently close to $O(n)$, which is the worst-case parametric updating complexity. Based on the new algorithm, the semiparametric estimator can be constructed within several minutes even when there are millions of data points. Moreover, when conducting the new algorithm, only roughly $O(\sqrt{n})$ data points have to be stored in the computer memory in each round of update, so it substaintially alleviates the memory requirement. More importantly, we show that the new estimator is $1/\sqrt{n}$-trivial with respect to the full-sample-based estimator constructed based on (2.2), implying that there will be no loss of estimation accuracy despite the substantial improvement in the computation speed.

The key technique adopted to relive the devastating computational burden is *subsampling*. Such technique is similar in spirit to but essentially different from the existing MBGD algorithm that we discussed before. According to our previous discussion, the heavy computational burden in update (2.2) is caused by full-sample-based update and Nadaraya-Watson kernel estimation. Even though stochastic optimization approach such as MBGD algorithm uses a small portion of data points to perform the update, full-sample-based Nadaraya-Watson kernel estimation still takes up a huge amount of computation time and memory requirement. Motivated by such observation, our new algorithm is fully subsample-based. In other words, in each round of update we will randomly draw a subsample $\mathfrak{I}_{B,k}$, and then use such subsample to both construct the Nadayara-Watson kernel estimators and perform the update. To be specific, in the k-th round of iteration we start with parameter $\boldsymbol{\beta}_{e,k}$, then consider the following Nadaraya-Watson kernel estimator of $G(z)$ constructed

[2]This is becuase, for example, when using MBGD algorithm to update the parameter, only B data points need to be accessed in each update, so only B data points have to be effectively stored in the computer memory.

based on the data points in subsample $\mathfrak{I}_{B,k}$,

$$\widehat{G}_n\left(z|\,\boldsymbol{\beta}_{e,k},\mathfrak{I}_{B,k},\underline{c}_f\right)=\frac{\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}K\left(\left(z-\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right)/h_n\right)y_j}{\left\{\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}K\left(\left(z-\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right)/h_n\right)\right\}\vee\underline{c}_f}, \tag{2.4}$$

where K and h_n are all similarly defined as before, and $\underline{c}_f>0$ is some sufficiently small constant. Such subsample-based kernel estimator is constructed as if we only observe the data points in the subsample $\mathfrak{I}_{B,k}$. Note that for both the numerator and denominator of (2.4), summation is only performed over B data points, so the computational time for evaluating the kernel estimator is of order $O(B)$. Given (2.4), we perform the following subsample-based update

$$\boldsymbol{\beta}_{e,k+1}=\boldsymbol{\beta}_{e,k}-\frac{\delta_k}{B}\sum_{i\in\mathfrak{I}_{B,k}}\frac{\partial L(\mathbf{X}_{e,i},y_i,\boldsymbol{\beta}_{e,k}|\widehat{G}_n\left(\cdot|\,\boldsymbol{\beta}_{e,k},\mathfrak{I}_{B,k},\underline{c}_f\right))}{\partial\boldsymbol{\beta}_e}, \tag{2.5}$$

Obviously, the computational time required for update in (2.5) is of order $O(B^2)$. According to our theoretical results, B can be chosen close to $\sqrt{n}$, indicating that the computational complexity of each update based on (2.5) is close to the order $O(n)$, which is almost linear in sample size n and is numerically feasible when n is large.

This paper contributes to the literature of semiparametric estimation of monotone index models. The existing estimation methods can be roughly classified into two categories: M-estimation approach and direct construction approach. For the first category, the estimator is obtained by optimizing some objective functions. The standing estimators include maximum score estimator (Manski, 1975, 1985; Horowitz, 1992), maximum rank correlation estimator (Han, 1987; Sherman, 1993; Cavanagh and Sherman, 1998; Fan et al., 2020), semiparametric least squares estimator (Härdle et al., 1993; Ichimura, 1993), semiparametric maximum likelihood estimator (Cosslett, 1983; Klein and Spady, 1993), and more recently KBGD and SBGD estimators (Khan et al., 2023). Apart from M-estimation, the second class of estimation methods features direct construction of the estimators, which includes average derivative estimator (Stoker, 1986; Powell et al., 1989; Horowitz and Härdle, 1996; Hristache et al., 2001), special regressor approach (Lewbel, 2000) and eigenvalue approach (Ahn et al., 2018). Unfortunately, almost all the existing methods can not be effectively applied when the sample size is extremely large. Moreover, apart from intensive computational burdens, there are many other crucial limitations that prohibit the use of existing methods in the empirical applications[3].

[3]For M-estimation, the objective functions are usually heavily discontinuous and/or non-convex with respect to the parameter. In this case, even looking for a local optimum is generally NP-Hard (Murty and Kabadi, 1987), let alone the

In principle, the general idea of our methodological development in this paper can be applied to any loss functions such as least squares (Ichimura, 1993) or maximum likelihood (Klein and Spady, 1993) loss functions , but to make our discussion more intuitive, we adopt the loss function used in Khan et al. (2023) (KLTY hereafter). KLTY consider a less commonly-used loss function whose derivative with respect to the parameter depends only on $G(\cdot)$ itself. They then propose an iterative algorithm that deviates from the conventional practice of gradient-based approach. In particular, instead of plugging the kernel estimator of $G(\cdot)$ into the loss function and then calculate the gradient of the plug-in loss function, they first calculate the gradient of the loss function as if $G(\cdot)$ is known. Then they replace the unknown $G(\cdot)$ in the gradient function with its Nadaraya-Watson kernel estimator or sieve estimator, and use the plug-in gradient to perform the update. They argue that the twisted algorithm ensures that the update effectively forms a contraction map for the parameter, and the resulting estimator based on such update is numerically robust to the choice of the initial guess. Similar to all of the semiparametric M-estimators such as those in Ichimura (1993) and Klein and Spady (1993), the key bottleneck of KLTY's method in the large n setup lies in the heavy computational burden, as we have discussed earlier. Indeed, when using kernel estimators as the replacement of the unknown function, n kernel estimators have to be evaluated in each update and each kernel estimator is constructed based on the full sample. This leads to a computational burden of order $O(n^2)$, making such method computational infeasible even for modest n, say, $n = 50000$.

Similar to the development from deterministic optimization to stochastic optimization, this paper develops a fully subsample-based estimation procedure based on KLTY's algorithm, where the nonparametric estimation as well as the update are all based on a random subsample from the full data set. Inheriting from the advantages of KLTY's method, our proposed method does not suffer from optimization issue and can be applied to the case where there are many covariates with mixture of both discrete and continuous ones. The key theoretical challenge arising from such methodological development lies in the sizable bias caused by subsample-based kernel estimation. In particular, when using subsamples to construct (2.4), subsample-based summations appear in both the denominator and the numerator, making (2.4) a biased estimator of the full-sample-based Nadaraya-Watson

global optimum. This makes the optimization procedure computationally infeasible (Khan et al., 2023). On the other side, the direct construction approach generally imposes more structure on the covariates. For example, the average derivative approach requires that the covariates are all continuous, so can not be directly applied to discrete covariates such as dummy variables. Moreover, the application of such method usually involves nonparametric estimation of the density functions or their partial derivative of some random variables conditional on the covariates. Such estimation becomes an intractable problem even when the number of covariates is modest. Although there have been some attempts to reduce the dimensionality of conditional density estimation (e.g., Hall et al. (2004)), the methods are still computationally-intensive, which may not be applicable in a data-rich environment, see Ouyang and Yang (2023) and references therein.

estimator. Such bias dampens the $1/\sqrt{B}$-convergence rate of the subsample-based estimator, making it converge at a much slower rate than standard subsample-based estimators[4]. We then proceed to decompose the bias. We find that the first-order bias has $1/\sqrt{n}$-trivial conditional mean (conditioned on the subsamples in the previous updates and the data set), while the second-order bias is uniformly $1/\sqrt{n}$-trivial as long as we update sufficiently many times. This motivates us to follow Polyak and Juditsky (1992) and use average to eliminate the first-order bias and accelerate the convergence rate. In particular, after some burn-in rounds of updates, all the estimators produced during the following updates are averaged. We show that as long as the numbers of burn-in and follow-up updates are both large enough, the averaged estimator will converge at $1/\sqrt{n}$ rate and is asymptotically normally distributed. Such a result demonstrates that our subsample-based method not only improves the computational speed, it also maintains the estimation accuracy on the same time.

Since the subsample-based estimator is asymptotically normally distributed after averaging, inference on the true parameter can be conducted if some consistent estimator of the asymptotic covariance matrix is available. Unfortunately, when sample size n is extremely large, estimating the covariance matrix based on the full sample also requires large amount of time because it involves evaluating a large number of nonparametric estimators. To faciliate the inference, we also propose a subsample-based estimator of the covariance matrix, which subtantially improves the computation speed. We show that the subsample-based estimator is a consistent estimator of the unknown covariance matrix, so the inference using such subsample-based estimator will be asymptotically valid.

The main contribution of this paper to the econometric and machine learning literature is that it proposes a novel computationally friendly algorithm that can be used to semiparametrically estimate the monotone index models even when the sample size is extremely large. Our new algorithm essentially generalizes the mini-batch estimation approach to the semiparametric setup. It can be easily applied when there are tens or hundreds of covariates and hundreds of thousands of or even millions of data points. Essentially, it bridges the gap between semiparametric estimation theories and empirical applications in the data-rich environment.

The remainder of the paper is arranged as follows. In section 2.2, we formally introduce the two-step fully subsample-based updating algorithm. In section 2.3, we develop the asymptotic properties of the proposed algorithm, and we also propose a subsample-based inference procedure. In section 3.4,

[4]For conventional MBGD estimators, $1/\sqrt{B}$-consistency is a standard convergence rate.

we illustrate the performance of new algorithm by conducting some Monte Carlo experiments. In section 3.5, we illustrate the empirical applicability of our new algorithm by analyzing three real world data sets. Finally, section 3.7 concludes. All the proofs of the lemmas and theorems are arranged to the Appendix.

2.1.1 Notations

For any real sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, write $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| = 0$, $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$, and $a_n \sim b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$. For any random sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, write $a_n = O_p(b_n)$ if for any $0 < \tau < 1$ there exist N and $C > 0$ such that $P\{|a_n/b_n| > C\} < \tau$ holds for all $n \geq N$, I write $a_n = o_p(b_n)$ if for any $C > 0$, $P(|a_n/b_n| > C) \to 0$. For any Borel set $A \subseteq \mathbb{R}^k$, denote its Lebesgue measure as $m(A)$. Denote I_p as the p-dimensional identity matrix. For any symmetric matrix A, we write $A \succ 0$ if A is positive definite, and $A \succeq 0$ if A is positive semi-definite. For any symmetric matrices A and B, I write $A \succ B$ if $A - B \succ 0$ and $A \succeq B$ if $A - B \succeq 0$. For any matrix A, denote $\sigma(A)$ as its singular value, and denote $\overline{\sigma}(A)$ and $\underline{\sigma}(A)$ as its largest and smallest singular value. For any symmetric matrix A, denote $\lambda(A)$ as its eigenvalue, and denote $\overline{\lambda}(A)$ and $\underline{\lambda}(A)$ as its largest and smallest eigenvalue. For any vector $\boldsymbol{x} = (x_1, \cdots, x_p)^{\mathrm{T}}$, denote its Euclidean norm as $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^{p} x_i^2}$. For any matrices $A = (a_{ij})_{n\times m}$, denote $\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$.

2.2 The Algorithm

To fix idea, throughout this paper we will focus on the following binary choice model

$$y = \mathbb{1}\left(X_0\beta_0^{\star} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^{\star} - u > 0\right), \tag{2.6}$$

where $\mathbb{1}(\cdot)$ is indicator function and u is the unobserved individual shock with CDF $G(\cdot)$. Binary choice model is a leading example of monotone index models, which has a wide range of applications in many areas such as economics, business, and biostatistics. We also emphasize that all of the conclusions obtained in the following paper can be applied to general class of monotone index models without any modifications of the algorithm.

We make the following assumptions regarding the data generating process (2.6) and the data set we

observe.

Assumption 2.1. *An i.i.d. data set* $\mathcal{D}_n = \{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^n$ *of sample size* n *is observed, where* y_i *is generated by* $y_i = \mathbb{1}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^\star - u_i > 0\right)$ *with unobserved shock* u_i *that is independent of* $\mathbf{X}_{e,i}$ *and has CDF* $G(\cdot)$.

Assumption 2.2. *(i)* $\mathcal{X}_e = [-1,1]^{p+1}$*; (ii)* $\mathcal{B}_e$ *is convex, and there exists some constant* $B_0 > 0$ *such that for any* $\boldsymbol{\beta}_e \in \mathcal{B}_e$, $|\beta_j| \leq B_0$ *for any* $0 \leq j \leq p$*; (iii) The CDF* G *has up to* $(D+1)$*-th order bounded derivatives.*

To make the illustration of the new algorithm more intuitive, we start with a special case where the CDF function $G(\cdot)$ is known. Following Agarwal et al. (2014) and Khan et al. (2023), we consider the loss function

$$L\left(\mathbf{X}_e, y, \boldsymbol{\beta}_e | G\right) = \int_{-A}^{\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e} G\left(z\right) dz - y\mathbf{X}_e^{\mathrm{T}}\boldsymbol{\beta}_e, \tag{2.7}$$

for some sufficiently large positive constant A. Khan et al. (2023) show that loss function (2.7) has many properties such as global minimization at true parameter $\boldsymbol{\beta}^\star$ and positive definite Hessian matrix with respect to $\boldsymbol{\beta}_e$. Based on the MBGD updating rule (2.3) and loss function (2.7), the MBGD estimator of $\boldsymbol{\beta}_e^\star$ is constructed based on the following iteration procedure:

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{B} \sum_{i \in \mathfrak{I}_{B,k}} \left(G\left(\mathbf{X}_{e,i}^{\mathrm{T}}\boldsymbol{\beta}_{e,k}\right) - y_i\right) \mathbf{X}_{e,i}, \tag{2.8}$$

where $\boldsymbol{\beta}_{e,1}$ is given, B is a positive integer and is the sbusample size. For each k, $\delta_k > 0$ is the learning rate, and

$$\mathfrak{I}_{B,k} = \{i_{k,1}, i_{k,2}, \cdots, i_{k,B}\} \tag{2.9}$$

is an index set that is randomly drawn from $\{1, 2, \cdots, n\}$ with replacement and is independent over k. In other words, under MBGD algorithm, in each iteration we randomly draw a subset of size B, and then update the estimator based on such subsample.

Now we turn to the case of semiparametric estimation, which is the main focus of this paper. To ensure identification, we set $\beta_0^\star$ to be 1, so the estimation target now is $\boldsymbol{\beta}^\star$. To simplify notation, denote the space of $\mathbf{X}$ as $\mathcal{X}$, and the corresponding parameter space of $\boldsymbol{\beta}$ as $\mathcal{B}$.

Remark 2.1. *Here we provide some discussion on the choice of the normalized covariate. The covariate whose coefficient is normalized to 1 must have nonzero and positive true coefficient. Since*

the true coefficient is unknown, we recommend choosing the covariate based on economic theories. However, there could be scenarios where the (unknown) actual coefficient has the opposite sign as to that implied by economic theories. So it's also recommend to conduct a preliminary estimation based on Logit or Probit to provide some additional insights. In particular, it's suggested to choose covariate whose coefficient is significantly different from zero. If the estimated coefficient is negative, then use the negative value of such covariate for estimation. Finally, it's also recommended using continuous variable as the normalized covariate.

Note that the MBGD algorithm (2.8) relies on the nonparametric component $G(\cdot)$ as a key input, which is unavailable in the current semiparametric setup. So the conventional MBGD algorithm is infeasible. To make the update feasible, a natural idea is to replace the unknown component with its nonparametric estimator. However, as we have discussed in section 2.1, if we use the conventional Nadaraya-Watson kernel estimator as did in Khan et al. (2023), even evaluating one estimator will take up $O(n)$ computational time, which leads to $O(Bn)$ computational time for a single update. To further relieve the computational burden, we propose to use the novel subsample-based kernel estimator (2.4). In particular, at the beginning of the k-th round of update, the initial point is given by $\boldsymbol{\beta}_k$. Then using the subsample-based kernel estimator of $G(z)$, consider the following update of $\boldsymbol{\beta}_k$,

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{B} \sum_{i \in \mathfrak{I}_{B,k}} \left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k, \mathfrak{I}_{B,k}, \underline{c}_f \right) - y_i \right) \mathbf{X}_i^{\phi}, \tag{2.10}$$

where $\mathbf{X}_i^{\phi} = \mathbf{X}_i \cdot \mathbb{1}(\mathbf{X}_{e,i} \in \mathcal{X}_e^{\phi})$, and $\mathcal{X}_e^{\phi} = \{\mathbf{X}_e \in \mathcal{X}_e : |X_j| \leq 1 - \phi, 0 \leq j \leq p\}$ for some $0 < \phi < 1$[5]. Since the above algorithm generalizes the conventional MBGD procedure to the semiparametric setup, we label the new algorithm the *Kernel-Based Mini-Batch Gradient Descent Algorithm* (KMBGD). The algorithm is summarized in algorithm 4.

We provide two more remarks.

Remark 2.2. *We provide some comparisons between the KMBGD algorithm and the KBGD algorithm proposed in KLTY. Basically, the latter algorithm is a full-sample-based algorithm; if we choose $\mathfrak{I}_{B,k} = \{1, \cdots, n\}$ for all k, then KMBGD degenerates to KBGD. For computational burden, we obviously have that KBGD has computational complexity of order $O(n^2)$ in each update, while the update of KMBGD has complexity of order $O\left(B^2\right)$. If we choose B close to $1/\sqrt{n}$, the computational complexity of KMBGD will be close to n, which is linear in the sample size and is roughly n times*

[5]Such truncation is basically used to improve the uniform convergence speed of kernel estimation. Similar method is applied in many research such as Ichimura (1993), Klein and Spady (1993), and Khan et al. (2023).

Algorithm 4: The KMBGD Estimator

input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^{n}$, sequence of learning rate $\{\delta_k\}_{k=1}^{\infty}$, initial guess $\boldsymbol{\beta}_1$, kernel function K, bandwidth h_n, subsample size B, number of iterations T, trimming parameter ϕ and $\underline{c}_f$

output: The KMBGD estimator $\widehat{\boldsymbol{\beta}}$

$k \leftarrow 1$;
while $k \leq T$ **do**
  Generate index set $\mathfrak{I}_{B,k}$;
  for $l \leftarrow 1$ **to** B **do**
    $$\widehat{G}\left(X_{0,i_{k,l}} + \mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k, \mathfrak{I}_{B,k}, \underline{c}_f\right) \leftarrow \frac{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,k}} K_{h_n}\left(X_{0,i_{k,l}}+\mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_j}{\left\{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,k}} K_{h_n}\left(X_{0,i_{k,l}}+\mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right\} \vee \underline{c}_f};$$
  $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{B}\sum_{i\in\mathfrak{I}_{B,k}}\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k, \mathfrak{I}_{B,k}, \underline{c}_f\right) - y_i\right)\mathbf{X}_i^{\phi}$;
  $k \leftarrow k+1$;
$\widehat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}_{T+1}$;

smaller than that of KBGD. This implies that when n is extremely large, KMBGD is a better option.

Remark 2.3. *Similar to the KBGD algorithm, our method is also iteration-based and does not rely on any optimization procedure, so it can be easily implemented when the number of the covaraites p is also large. In other words, the KMBGD estimator applies to the scenario where both n and p are large. However, since in this paper we mainly focus on the scenario where the sample size n is extremely large, in my following theoretical analysis we will take p as being fixed.*

2.3 Asymptotic Properties of KMBGD Estimator

In this section, we formally develop the statistical properties of the proposed KMBGD estimator. Under some regularity conditions, we first show that as long as we update the parameter sufficiently many times, the KMBGD estimator is consistent. However, the convergence rate is slower than $1/\sqrt{n}$ if we choose $B \ll n$. Indeed, such rate is even slower than $1/\sqrt{B}$, which is the convergence rate of general mini-batch estimators. Then we will show that although KMBGD estimator itself converges at a slow rate, we can conduct averages across all the estimators produced during updates to accelerate the convergence rate. In particular, we show that if we properly choose subsample size, bandwidth prameter, order of kernel function, and number of iterations, the average estimator obtains $1/\sqrt{n}$-consistency.

Before we illustrate the main results, we first introduce some notations. Let $f_e(\mathbf{X}_e)$ and $f(\mathbf{X})$

denote the joint density of $\mathbf{X}_e$ and $\mathbf{X}$[6]. Define $z(\mathbf{X}_e, \boldsymbol{\beta}) = X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. Let $f_{\mathbf{X}|z}(\mathbf{X}|\, z, \boldsymbol{\beta})$ be the conditional density of $\mathbf{X}$ given $z(\mathbf{X}_e, \boldsymbol{\beta}) = z$ and $\boldsymbol{\beta}$. Define

$$
\begin{aligned}
W\left(\mathbf{X}_e, \widetilde{\mathbf{X}}_e, \boldsymbol{\beta}\right) &= G'\left(z(\mathbf{X}_e, \boldsymbol{\beta}^\star) + \left(\mathbf{X} - \widetilde{\mathbf{X}}\right)^{\mathrm{T}} \Delta\boldsymbol{\beta}\right) f_{\mathbf{X}|z}\left(\widetilde{\mathbf{X}}\middle|\, z(\mathbf{X}_e, \boldsymbol{\beta}), \boldsymbol{\beta}\right), \\
V\left(\mathbf{X}_e, \widetilde{\mathbf{X}}_e, \boldsymbol{\beta}\right) &= \left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{X}\widetilde{\mathbf{X}}^{\mathrm{T}}\right) W\left(\mathbf{X}_e, \widetilde{\mathbf{X}}_e, \boldsymbol{\beta}\right), \\
\Lambda_\phi(\boldsymbol{\beta}) &= \mathbb{E}\left[\mathbb{1}_i^\phi \cdot \int_{\mathcal{X}} V(\mathbf{X}_{e,i}, \mathbf{X}_e, \boldsymbol{\beta})\, d\mathbf{X}\right].
\end{aligned}
$$

The following additional technical assumptions are imposed.

Assumption 2.3. *The kernel function $K(\cdot)$ satisfies: (i) K is bounded and twice continuously differentiable with bounded first and second derivatives, and the second derivative satisfies Lipschitz condition on the whole real line; (ii) $\int K(s)\, ds = 1$; (iii) $\int s^\upsilon K(s)\, du = 0$ for $1 \le \upsilon \le D-1$ and $\int u^D K(u)\, du \neq 0$; (iv) $K(s) = 0$ for $|s| > 1$.*

Assumption 2.4. *(i) There exists some constant $\zeta > 1$ such that $\zeta^{-1} \le f_e(\mathbf{X}_e) \le \zeta$ holds for all $\mathbf{X}_e \in \mathcal{X}_e$; (ii) $f_e(\mathbf{X}_e)$ has up to $(D+1)$-th order bounded derivatives.*

Assumption 2.5. *There hold $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \overline{\lambda}\left(\Lambda_0(\boldsymbol{\beta}) + \Lambda_0^{\mathrm{T}}(\boldsymbol{\beta})\right) \le \overline{\lambda}_\Lambda < \infty$, and $\inf_{\boldsymbol{\beta}\in\mathcal{B}} \underline{\lambda}\left(\Lambda_0(\boldsymbol{\beta}) + \Lambda_0^{\mathrm{T}}(\boldsymbol{\beta})\right) \ge \underline{\lambda}_\Lambda > 0$.*

All the above assumptions are also imposed in KLTY. Based on the above assumptions, now we formally study the statistical properties of the iterative estimator $\boldsymbol{\beta}_k$ based on iteration (2.4) and (2.10). We first introduce some further notations. Let P denote the probability measure of the data set $\mathcal{D}_n$. Let $\mathbb{P}^*$ be the probability measure corresponding to random variables $\{\mathfrak{I}_{B,k}\}_{k=1}^\infty$ and $\mathbb{P}_k^*$ be probability measure corresponding to $\{\mathfrak{I}_{B,k'}\}_{k'\ge k}^\infty$ conditional on the observation of $\{\mathfrak{I}_{B,k'}\}_{k'=1}^{k-1}$ for $k \ge 2$ and $\mathbb{P}_1^* = \mathbb{P}^*$. Let $\mathbb{E}^*$ and $\mathbb{E}_k^*$ be the expectation with respect to $\mathbb{P}^*$ and $\mathbb{P}_k^*$. Finally, let $\mathbb{P}$ be the probability measure of $\{\mathcal{D}_n, \mathfrak{I}_{B,1}, \mathfrak{I}_{B,2}, \cdots\}$, where $\mathcal{D}_n$ is the data set.

Recall that the Nadaraya-Watson kernel estimator for $\mathbb{E}\left(y|\, X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = z\right)$ based on the full data is given by $\widehat{G}(z|\,\boldsymbol{\beta})$. For any $\boldsymbol{\beta} \in \mathcal{B}$, define $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^\star$. We obviously have the following

[6]By assuming $\mathbf{X}_e$ has joint density function, we require that $\mathbf{X}_e$ is continuous, which facilitates our following discussion. However, we point out that our analysis can be trivially extended to the case where there are some discrete covariates, see KLTY.

decomposition for the KMBGD update (2.10),

$$\begin{aligned}\Delta\boldsymbol{\beta}_{k+1} = \Delta\boldsymbol{\beta}_k &- \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\middle|\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi} \\ &\underbrace{-\delta_k\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}\left(\widehat{G}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\middle|\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi}-\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\middle|\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi}}_{\pi_{1,n,k}} \\ &\underbrace{-\delta_k\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}\left(\widehat{G}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\middle|\boldsymbol{\beta}_k,\mathfrak{I}_{B,k},\underline{c}_f\right)-\widehat{G}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k\middle|\boldsymbol{\beta}_k\right)\right)\mathbf{X}_i^{\phi}}_{\pi_{2,n,k}}. \end{aligned} \tag{2.11}$$

It's not difficult to see that if $\pi_{1,n,k}=\pi_{2,n,k}=0$, then (2.11) degenerates to the full-sample-based KBGD algorithm. Indeed, $\pi_{1,n,k}$ describes the randomness caused by updating using only a subset of the data, whereas $\pi_{2,n,k}$ describes the randomness caused by performing nonparametric kernel estimation using only a subset of the data points. Essentially, $\pi_{1,n,k}$ is shared by all the mini-batch estimators, while $\pi_{2,n,k}$ is specific to the semiparametric setup we consider in this paper. We have the following lemma describing the properties of $\pi_{1,n,k}$ and $\pi_{2,n,k}$.

Lemma 2.1. *Suppose that Assumption 2.1–Assumption 2.5 hold with $D\geq 4$. Suppose also that $\underline{c}_f$ is chosen such that $\inf_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}} f_Z\left(z\middle|\boldsymbol{\beta}\right)\geq 3\underline{c}_f$. If $\boldsymbol{\beta}_k$ is update based on (2.4) and (2.10), We have that*

$$P\left[\sup_{k\geq 1}\mathbb{E}^*\left(\left\|\pi_{1,n,k}\right\|^2\right)\leq\frac{C}{B}\right]\to 1,$$

and

$$P\left[\sup_{k\geq 1}\mathbb{E}^*\left(\left\|\pi_{2,n,k}\right\|^2\right)\leq\frac{C\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right]\to 1,$$

for some C that does not depend on n, B, h_n, and k.

Lemma 2.1 immediately yields the following result.

Theorem 2.1. *Suppose that Assumption 2.1–Assumption 2.5 hold with $D\geq 4$. Suppose also that $\underline{c}_f$ is chosen such that $\inf_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}} f_Z\left(z\middle|\boldsymbol{\beta}\right)\geq 3\underline{c}_f$. Suppose moreover that $\delta_k=\delta<\min\{1/\left(2\underline{\lambda}_{\Lambda}\right),1/\left(4p^2\left\|G'\right\|_{\infty}\right)\}$, $\phi<\delta\underline{\lambda}_{\Lambda}/\left(16p^2\left\|G'\right\|_{\infty}\zeta\right)$, h_n is chosen such that $h_n n^{1/2D}\to 0$ and $h_n n^{1/6}/\log^{1/3}(n)\to\infty$. If $\boldsymbol{\beta}_k$ is update based on (2.4) and (2.10), define*

$$k_n=\left\lceil\frac{\log\left(h_n^{2D}+\sqrt{\log\left(Bh_n^{-2}\right)/Bh_n^2}\right)-\log\left(\sqrt{\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_1\right\|^2}\right)}{\log\left(1-\delta\underline{\lambda}_{\Lambda}/8\right)}\right\rceil,$$

we have that

$$\sup_{k\geq k_n+1}\mathbb{E}^*\left(\|\Delta\boldsymbol{\beta}_k\|^2\right)=O_p\left(h_n^{2D}+\frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right).$$

According to Theorem 2.1, if we choose $B\ll n$ to improve computational speed, the upper bounded on the estimation error $\mathbb{E}^*\left(\|\Delta\boldsymbol{\beta}_k\|\right)$ will be of rate slower than $n^{-1/2}$ even when the order of the kernel function is large. The slower convergence rate is a common feature of all the mini-batch estimators. Indeed, the mini-batch estimators converge at the rate $1/\sqrt{B}$ at best, see, for example, Lemma 2 in Forneron (2022). However, different from the conventional mini-batch estimator, my KMBGD estimators are guaranteed to converge no faster than $\sqrt{\log(n)/Bh_n^2}$. If I choose $B=1/\sqrt{n}$ and $h_n=n^{-1/6}$, then the convergence rate would be $\sqrt{\log(n)}n^{-1/12}$, which is much slower than $1/\sqrt{B}=n^{-1/4}$.

The slower convergence rate of the KMBGD estimator is mainly due to the fact that we use subsamples to construct the kernel estimator. In this case, the subsample-based gradient is no longer an unbiased estimator (conditional on the previous subsamples) of the full-sample-based gradient, that is, $\mathbb{E}^*(\pi_{2,n,k})\neq 0$. The bias makes the convergence rate of KMBGD estimator slower than $1/\sqrt{B}$. However, surprisingly, in the following we will show that if we appropriately choose the kernel function and bandwidth parameter, even with $B\ll n$, we can still obtain $1/\sqrt{n}$ by following Polyak and Juditsky (1992) and conducting average across KMBGD estimators produced during iterations.

To formally show the above results, we first further decompose the KMBGD dynamics. To ease our following exposition, for any z and $\boldsymbol{\beta}$ denote $A_{n,y}\left(z,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^n K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)y_i$, $A_{n,1}\left(z,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^n K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $A_{n,y}\left(z,\boldsymbol{\beta}|\,\mathfrak{I}_{B,k}\right)=\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)y_i$, and $A_{n,1}\left(z,\boldsymbol{\beta}|\,\mathfrak{I}_{B,k}\right)=\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$. We have the following lemma.

Lemma 2.2. *Suppose that all the assumptions and conditions in Theorem 2.1 hold. Suppose moreover that* $B\cdot\min\{h_n^6/\log^2(n),h_n^2/(\sqrt{n}\log(n))\}\to\infty$. *Define* $\boldsymbol{\xi}_n^{\phi}=\frac{1}{n}\sum_{i=1}^n(\widehat{G}\left(z_i^{\star}|\,\boldsymbol{\beta}^{\star}\right)-y_i)\mathbf{X}_i^{\phi}$, *where* $z_i^{\star}=z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}^{\star}\right)$. *Also define* $z_{i,k}=z(\mathbf{X}_{e,i},\boldsymbol{\beta}_k)$. *If* $\boldsymbol{\beta}_k$ *is update based on (2.4) and (2.10),*

we have that

$$\begin{aligned}\Delta\boldsymbol{\beta}_{k+1} = &\left(I_p - \delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)\Delta\boldsymbol{\beta}_k - \delta\boldsymbol{\xi}_n^\phi + \delta\Omega_k^\phi \\ &\underbrace{- \delta\frac{1}{B}\sum_{i\in\Im_{B,k}}\left(\widehat{G}\left(z_{i,k}|\,\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^\phi - \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}|\,\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^\phi}_{\varrho_{1,n,k}} \\ &\underbrace{- \delta\frac{1}{B}\sum_{i\in\Im_{B,k}}\frac{\mathbf{X}_i^\phi}{A_{n,1}\left(z_{i,k},\boldsymbol{\beta}_k\right)}\cdot\left(A_{n,y}\left(z_{i,k},\boldsymbol{\beta}_k|\,\Im_{B,k}\right) - A_{n,y}\left(z_{i,k},\boldsymbol{\beta}_k\right)\right)}_{\varrho_{2,n,k}} \\ &\underbrace{+ \delta\frac{1}{B}\sum_{i\in\Im_{B,k}}\frac{A_{n,y}\left(z_{i,k},\boldsymbol{\beta}_k\right)\mathbf{X}_i^\phi}{A_{n,1}^2\left(z_{i,k},\boldsymbol{\beta}_k\right)}\cdot\left(A_{n,1}\left(z_{i,k},\boldsymbol{\beta}_k|\,\Im_{B,k}\right) - A_{n,1}\left(z_{i,k},\boldsymbol{\beta}_k\right)\right),}_{\varrho_{3,n,k}}\end{aligned}$$

where $\sup_{k\geq k_n+1}\mathbb{E}^*\left\|\Omega_k^\phi\right\| = o_p\left(n^{-1/2}\right)$.

We now provide some intuitive discussion for Lemma 2.2. Basically, if there are no noise terms $\varrho_{1,n,k}$, $\varrho_{2,n,k}$, and $\varrho_{3,n,k}$, then the dynamics of $\Delta\boldsymbol{\beta}_k$ simply degenerate to the full-sample-based KBGD algorithm in KLTY as implied in Lemma 2.3 in Appendix. However, since we use subsamples to perform the update, additional noises due to subsampling are introduced into the update and these noises are captured by the above three terms. Basically, $\varrho_{1,n,k}$ describes the impacts of using subsamples instead of full sample to perform the update. Such error is shared by all the mini-batch-based methods. While the remaining two terms $\varrho_{2,n,k}$ and $\varrho_{3,n,k}$ describe the impacts of using subsamples instead of full sample to construct the Nadaraya-Watson kernel estimator, so are specific to my algorithm only. Simple calculation leads to $\mathbb{E}^*\left(\varrho_{1,n,k}\right) = 0$, $\mathbb{E}^*\left(\varrho_{2,n,k}\right) = O_p\left(1/Bh_n\right)$, and $\mathbb{E}^*\left(\varrho_{3,n,k}\right) = O_p\left(1/Bh_n\right)$ uniformly with respect to k. The above implies that for k sufficiently large, the first-order difference between KBGD and KMBGD estimators almost constitute a martingale difference sequence. By "almost" we mean that the conditional expectation is of order $O_p(1/Bh_n)$, which can be made $n^{-1/2}$-trivial if we choose $B \gg n^{1/2}h_n^{-1}$.

Lemma 2.2 implies that although the KMBGD estimator itself does not obtain $1/\sqrt{n}$-consistency due to noises caused by subsample-based kernel estimation and update, we can follow Polyak and Juditsky (1992) to conduct average across the estimators produced during iterations to eliminate these noises. Similar to the conventional mini-batch gradient estimator, the resulting estimator will be $1/\sqrt{n}$-consistent as long as we choose B that diverges at some rate. In particular, let k^* be the number of burn-in iterations and T be the number of follow-up iterations. The averaged KMBGD

estimator (AKMBGD) is defined as follws,

$$\overline{\boldsymbol{\beta}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}_{k^*+t}. \tag{2.12}$$

We summarize the algorithm in algorithm 5.

Algorithm 5: The AKMBGD Estimator

input : Data set $\{(\mathbf{X}_{e,i}, y_i)\}_{i=1}^{n}$, sequence of learning rate $\{\delta_k\}_{k=1}^{\infty}$, initial guess $\boldsymbol{\beta}_1$, kernel function K, bandwidth h_n, subsample size B, number of burn-in iterations k^*, number of follow-up iterations T, trimming parameter ϕ and $\underline{c}_f$

output: The AKMBGD estimator $\overline{\boldsymbol{\beta}}$

$k \leftarrow 1$;
while $k \leq k^* + T$ **do**
  Generate index set $\mathfrak{I}_{B,k}$;
  for $l \leftarrow 1$ **to** B **do**
    $\widehat{G}\left(X_{0,i_{k,l}} + \mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k, \mathfrak{I}_{B,k}, \underline{c}_f\right) \leftarrow \frac{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,k}} K_{h_n}\left(X_{0,i_{k,l}}+\mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right) y_j}{\left\{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,k}} K_{h_n}\left(X_{0,i_{k,l}}+\mathbf{X}_{i_{k,l}}^{\mathrm{T}}\boldsymbol{\beta}_k - X_{0,j} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}_k\right)\right\}\vee\underline{c}_f}$;
  $\boldsymbol{\beta}_{k+1} \leftarrow \boldsymbol{\beta}_k - \frac{\delta_k}{B}\sum_{i\in\mathfrak{I}_{B,k}}\left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^{\phi}$;
  $k \leftarrow k + 1$;
$\overline{\boldsymbol{\beta}} \leftarrow \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}_{k^*+t}$;

Now we provide the theoretical properties of the AKMBGD estimator.

Theorem 2.2. *Suppose that all the assumptions and conditions in Theorem 2.1 hold. Suppose moreover that* $B\cdot\min\{h_n^6/\log^2(n), h_n^2/(n^{1/2}\log(n))\} \to \infty$. *Let* $k^* = k_n + [-\log(n)/\log(1-\delta\underline{\lambda}_{\Lambda}/8)]$. *If* $\boldsymbol{\beta}_k$ *is update based on (2.4) and (2.10), for any* $T \geq 1$, *we have that*

$$\Delta\overline{\boldsymbol{\beta}} = -\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\boldsymbol{\xi}_n^{\phi} + O_{\mathbb{P}}\left(\frac{1}{\sqrt{Bh_n^2T}} + \frac{\log^{1/4}(n)}{Bh_n}\right).$$

If T *is chosen such that* $Bh_n^2Tn^{-1} \to \infty$, *we have that*

$$\sqrt{n}\Delta\overline{\boldsymbol{\beta}} \to_d \mathcal{N}\left(0, \Sigma_{\boldsymbol{\beta}}^{\phi}\right),$$

where $\Sigma_{\boldsymbol{\beta}}^{\phi} = \Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\Sigma_{\boldsymbol{\xi}}^{\phi}\left(\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}$ *and*

$$\Sigma_{\boldsymbol{\xi}}^{\phi} = \mathbb{E}\left[\left(1 - G\left(z_i^{\star}\right)\right)G\left(z_i^{\star}\right)\left(\mathbf{X}_i^{\phi} - \mathbb{E}\left(\mathbf{X}_i^{\phi}\middle| z_i^{\star}\right)\right)\left(\mathbf{X}_i^{\phi} - \mathbb{E}\left(\mathbf{X}_i^{\phi}\middle| z_i^{\star}\right)\right)^{\mathrm{T}}\right].$$

Theorem 2.2 is the key result of this paper. It demonstrates that even though we only use a random

subsample whose size is substaintially smaller than the full sample size to conduct kernel estimation and perform update in each round of iteration, the average of estimators produced during iterations will be equivalent to the full-sample estimator up to some small order terms. The small order terms will be uniformly $1/\sqrt{n}$-trivial as long as we choose $B \gg \max\{\log^2(n)h_n^{-6}, \sqrt{n}\log(n)h_n^{-2}\}$ and $T \gg nB^{-1}h_n^{-2}$. This implies that as long as we choose kernel function properly, the KMBGD estimator will be as efficient as the one based on the full sample, despite the fact that we only use a much smaller subsample to perform the update in each round.

Theorem 2.2 also suggests that the computational speed of each update can be improved by appropriately choosing the kernel function. In particular, since h_n must satisfy $h_n \ll n^{-1/2D}$ according to the conditions required in the theorem, then $B \gg \max\{n^{3/D}\log^2(n), n^{1/2+1/D}\log(n)\}$ must hold, so the computational complexity will be of order at least $O(\max\{n^{6/D}\log^4(n), n^{1+2/D}\log^2(n)\})$. Obviously, to improve the computational speed, we can choose a high-order kernel function. For example, if we choose a 8-th order kernel, the computational complexity is of order $O(n^{5/4}\log^2(n))$; if we choose a 12-th order kernel, the computational complexity is of order $O(n^{7/6}\log^2(n))$. If we can choose sufficiently large D, then the computational complexity is lower bounded by $n\log^2(n)$, which is almost the linear rate $O(n)$.

We finally discuss the total computational time of KBGD and KMBGD estimation. Suppose k^* updates are necessary to eliminate the impacts of the initial guess, then the full-sample-based KBGD algorithm requires $O(k^*n^2)$ computational time in total, while the KMBGD algorithms requires $O(k^*B^2 + B^2T)$. Since Theorem 2.2 requires that $T \gg nB^{-1}h_n^{-2}$, then the total computational time of KMBGD will be at least $O(k^*B^2 + nBh_n^{-2})$. If we choose $B \gg \sqrt{n}h_n^{-2}\log n$ and $h_n \ll n^{-1/2D}$, then $k^*B^2 + nBh_n^{-2} \gg k^*n^{1+2/D}\log^2(n) + n^{3/2+2/D}$. So the upper bound on the ratio between the total computational time of KBGD and KMBGD is of order

$$n^{1-2/D}\log^{-2}(n) + k^*n^{1/2-2/D}.$$

Obviously, when $D \geq 6$, the above ratio diverage at rate $n^{2/3} + k^*n^{1/6}$. More crucially, the above rate will be large when k^*, the number of burn-in updates, is large, which will often be the case when the number of covariates is large and $\underline{\Lambda}/\overline{\Lambda}$ is small,

Remark 2.4. *We provide some guidance on the applications of the KMBGD estimation. We recommend standardizing all the covariates*[7] *before estimation to improve the numerical performance. Re-*

[7]For any covariate w, the standardized covariate is given by $(w - \overline{w})/\sigma_w$, where $\overline{w}$ is the sample mean of w and

garding the choice of the tuning parameter, we recommend choosing $\delta_k = 1$ for all k in the first place, and if the iteration does not converge (around a fixed point), gradually shrink it towards zero, say, try $\delta_k = 0.1$ and $\delta_k = 0.01$. For the choice of B, we recommend choosing $B = \max\{3000, \sqrt{n}h_n^{-2}\log(n)\}$. For the stopping rule, we recommend updating until the mean of the estimators produced during iterations is stable. For example, let T and gap be two positive integers. First update the parameter $T + gap$ rounds. Then for each $k > T + gap$, compare two average estimators $\frac{1}{T}\sum_{j=1}^{T}\boldsymbol{\beta}_{k-j}$ and $\frac{1}{T}\sum_{j=1}^{T}\boldsymbol{\beta}_{k-j-gap}$. If the maximum distance between arguments of the above two estimators is smaller than some given tolerance ϱ, then stop and use the average of last $T + gap$ estimators as the final estimator. For another example, we can choose some pre-specified numbers of burn-in and follow-up updates, as long as both are sufficiently large.

We finally discuss the inference-related issues when the sample size n is large. According to Theorem 2.2, the AKMBGD estimator is asymptotically normally distributed, so inference on the true parameter $\boldsymbol{\beta}^\star$ can be conducted if we can consistently estimate the asymptotic covariance matrix $\Sigma_{\boldsymbol{\beta}}^{\phi}$. In their paper, KLTY provide a consistent estimator for the covariance matrix based on the full sample. However, to construct such estimator, we need to construct nonparametric estimators for conditional expectation $\mathbb{E}(\mathbf{X}_i^{\phi}|z_i^{\star})$ for each i, which may cost large amount of time when both n and p are large.

To solve the above inference issue in the large n scenario, this section provides a subsample-based estimator for the covariance matrix. Let $\{\mathfrak{I}_{B,r}\}_{r=1}^{R}$ be a sequence of random index sets defined in (2.9). For each $1 \le r \le R$, define

$$\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi,r} = \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,r}}\left(\widehat{G}_i^r\left(1-\widehat{G}_i^r\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}^r\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}^r\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)^{\mathrm{T}}\right),$$

and

$$\widehat{\Lambda}_{\phi}^r\left(\overline{\boldsymbol{\beta}}\right) = \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,r}}\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\overline{\boldsymbol{\beta}}\right)\middle|\overline{\boldsymbol{\beta}},\mathfrak{I}_{B,r},\overline{c}_f\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}},$$

where

$$\widehat{G}_i^r = \frac{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,r}}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)y_j}{\left\{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,r}}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)\right\}\vee\overline{c}_f},\ \widehat{\mathbb{E}}^r\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right) = \frac{\frac{1}{B}\sum_{j\in\mathfrak{I}_{B,r}}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)\mathbf{X}_j^{\phi}}{\left\{\frac{1}{B}\sum_{j=\in\mathfrak{I}_{B,r}}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)\right\}\vee\overline{c}_f},$$

σ_w is the sample standard deviation.

and $\widehat{z}_i = X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\overline{\boldsymbol{\beta}}$. Also define

$$\widetilde{\Sigma}_{\boldsymbol{\beta}}^{\phi} = \left(\frac{1}{R}\sum_{i=1}^{R}\widehat{\Lambda}_{\phi}^{r}\left(\overline{\boldsymbol{\beta}}\right)\right)^{-1}\left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi,r}\right)\left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\Lambda}_{\phi}^{r\mathrm{T}}\left(\overline{\boldsymbol{\beta}}\right)\right)^{-1}. \tag{2.13}$$

Then we have the following result.

Theorem 2.3. *Suppose that all the assumptions and conditions in Theorem 2.2 hold. If $Bh_n^2 \to \infty$, we have that*

$$\left\|\mathbb{P}^{*}\lim_{R\to\infty}\widetilde{\Sigma}_{\boldsymbol{\beta}}^{\phi} - \Sigma_{\boldsymbol{\beta}}^{\phi}\right\| \to_{\mathbb{P}} 0,$$

where $\mathbb{P}^{*}$ and $\mathbb{P}$ are defined in section 2.3. Moreover,

$$\widetilde{\Sigma}_{\boldsymbol{\beta}}^{\phi-1/2}\sqrt{n}\Delta\overline{\boldsymbol{\beta}} \to_{d} \mathcal{N}(0, I_p).$$

2.4 Monte Carlo Experiments

This section conducts some Monte Carlo experiments to evaluate the finite-sample performance of the KMBGD algorithm. We consider the following data generating process

$$y_i = \mathbb{1}\left(X_{0,i} + \beta_1^{\star}X_{1,i} + \cdots + \beta_9^{\star}X_{9,i} - u_i > 0\right), 1 \le i \le n, \tag{2.14}$$

where n is the sample size, and $(X_{0,i}, \cdots, X_{9,i}, u_i)$ is iid over i. For all $1 \le i \le n$, $X_{0,i} \sim \mathcal{N}(0,1)$, $X_{1,i} \sim$ Bernoulli(1/2), $X_{2,i} \sim$ Poisson(2), and $X_{j,i} \sim (\chi^2(1) - 1)/\sqrt{2}$ for $3 \le j \le 9$. So we have a mixture of both continuous and discrete covariates. Moreover, $X_{j,i}$ is independent over j for each i. u_i is the random error with cumulative distribution function $G(u)$, which is independent of the covariates. We consider four setups of error distrubtion: Cauchy, $t(4)$, $\chi^2(3)$, and $\mathcal{N}(0,1)$. We set the true parameter vector as $\boldsymbol{\beta}^{\star} = (1, 1, 0.5, 2, 5, -0.5, -1, -2, -5)^{\mathrm{T}}$. Finally, in the following simulations, whenever we conduct kernel estimation, we use eighth-order Epanechnikov kernel to construct the Nadaraya-Watson estimator, where the kernel function is given by $K(u) = 16.15(1 - u^2)(0.1667 - 1.5u^2 + 3.3u^4 - 2.043u^6) \cdot \mathbb{1}(|u| \le 1)$.

Table 2.1: Finite Sample Performance of Kernel-Based Estimators

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
					$u_i \sim$ Cauchy					
	Bias	0.0051	0.0010	0.0016	0.0042	0.0088	0.0003	0.0015	0.0041	0.0100
$n = 50000$	RMSE	0.0533	0.0314	0.0309	0.0610	0.1305	0.0258	0.0326	0.0549	0.1222
	CR	0.9570	0.9520	0.9490	0.9660	0.9660	0.9590	0.9580	0.9550	0.9670
	Bias	0.0006	0.0007	0.0003	0.0004	0.0016	0.0003	0.0009	0.0012	0.0036
$n = 100000$	RMSE	0.0366	0.0208	0.0206	0.0425	0.0924	0.0173	0.0229	0.0379	0.0879
	CR	0.9580	0.9590	0.9530	0.9490	0.9540	0.9640	0.9540	0.9570	0.9480
					$u_i \sim t(4)$					
	Bias	0.0023	0.0003	0.0000	0.0014	0.0019	0.0002	0.0004	0.0011	0.0019
$n = 50000$	RMSE	0.0362	0.0201	0.0187	0.0397	0.0869	0.0169	0.0213	0.0357	0.0805
	CR	0.9420	0.9490	0.9470	0.9600	0.9450	0.9430	0.9520	0.9470	0.9530
	Bias	0.0001	0.0001	0.0000	0.0004	0.0003	0.0003	0.0001	0.0005	0.0011
$n = 100000$	RMSE	0.0245	0.0138	0.0135	0.0273	0.0588	0.0115	0.0148	0.0248	0.0559
	CR	0.9490	0.9470	0.9490	0.9470	0.9600	0.9540	0.9580	0.9530	0.9650
					$u_i \sim \chi^2(3)$					
	Bias	0.0018	0.0015	0.0005	0.0008	0.0033	0.0001	0.0007	0.0001	0.0038
$n = 50000$	RMSE	0.0429	0.0246	0.0225	0.0482	0.1076	0.0217	0.0289	0.0458	0.1077
	CR	0.9590	0.9400	0.9490	0.9430	0.9380	0.9520	0.9450	0.9410	0.9420
	Bias	0.0001	0.0000	0.0002	0.0008	0.0020	0.0002	0.0001	0.0004	0.0002
$n = 100000$	RMSE	0.0301	0.0163	0.0159	0.0322	0.0718	0.0149	0.0197	0.0300	0.0707
	CR	0.9480	0.9540	0.9550	0.9490	0.9550	0.9620	0.9520	0.9650	0.9550
					$u_i \sim \mathcal{N}(0,1)$					
	Bias	0.0006	0.0001	0.0001	0.0004	0.0007	0.0004	0.0005	0.0006	0.0021
$n = 50000$	RMSE	0.0315	0.0166	0.0167	0.0347	0.0762	0.0145	0.0182	0.0306	0.0712
	CR	0.9500	0.9580	0.9570	0.9540	0.9500	0.9480	0.9590	0.9470	0.9420
	Bias	0.0001	0.0003	0.0008	0.0012	0.0007	0.0002	0.0002	0.0000	0.0000
$n = 100000$	RMSE	0.0214	0.0120	0.0119	0.0247	0.0534	0.0104	0.0134	0.0219	0.0506
	CR	0.9510	0.9590	0.9430	0.9480	0.9540	0.9510	0.9410	0.9560	0.9590

2.4.1 Finite-Sample Performance

We first study the finite sample performance of our AKMBGD estimator. We consider two sample sizes: $n = 50000$, and $n = 100000$. We report the bias, root mean squared error (RMSE), and coverage rate of AKMBGD estimators for $\beta_1^\star$ to $\beta_9^\star$. Suppose that the simulation is repeated R times, in the r-th round the estimator of $\beta_j^\star$ is denoted as $\widehat{\beta}_j^r$. Then the bias and RMSE of $\beta_j^\star$ is defined by

$$\text{Bias} = \left| \frac{1}{R} \sum_{r=1}^{R} \widehat{\beta}_j^r - \beta_j^\star \right|, \quad \text{RMSE} = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\widehat{\beta}_j^r - \beta_j^\star \right)^2}.$$

We consider nominal coverage rate 0.95, so the actual coverage rate is given by

$$\text{CR} = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1} \left(\widehat{\beta}_j^r - 1.96 \widehat{\sigma}_j^r \leq \beta_j^\star \leq \widehat{\beta}_j^r + 1.96 \widehat{\sigma}_j^r \right),$$

where $\widehat{\sigma}_j^r$ is the subsample-based estimator of the variance of $\widehat{\beta}_j^r$.

The learning rate is chosen as $\gamma_k = 1$ for all k. The bandwidth used in the k-th round of update is $h_n = c_k \cdot h_n^{-1/10}$, where $c_k = \text{std}\,(z_{i,k})$ and $z_{i,k} = X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k$. The initial guess is chosen as the Logit estimator. When constructing the AKMBGD estimator, I first run 2000 burn-in updates. Then the stopping rule is chosen as that in Remark 2.4 with $T = 10000$, $gap = 1000$, and $\varrho = 0.001$. The subsample size B is chosen as 3000 for both estimation and inference. Finally, when conducting inference, i randomly draw 200 subsamples to construct the variance estimator.

The simulation results are reported in Table 2.1. It can be seen that the AKMBGD estimators have small bias, whose RMSE decreases with sample size almost at rate $\sqrt{n}$. Moreover, the confidence interval constructed based on the subsample-based variance has actual coverage rate that is quite close to the nominal rate 0.95. This demonstrates that the AKMBGD estimators and subsample-based variance estimator have great finite-sample performance.

2.4.2 Computational Efficiency

This subsection formally compares the computational efficiency of several gradient-based estimators for semiparametric montone index models. In particular, I compare KMBGD estimator with the KBGD and SBGD estimators proposed by Khan et al. (2023).

I first compare the updating speed of each algorithm under different setups of sample sizes. In partic-

Table 2.2: Comparing Updating Speed

Sample Size	Method	KBGD	SBGD	KMBGD
$n = 2500$	Unparalleled	0.0475	0.0003	0.0081
	Parallel	0.0412	–	0.0321
$n = 5000$	Unparalleled	0.2009	0.0004	0.0078
	Parallel	0.0669	–	0.0292
$n = 10000$	Unparalleled	0.8335	0.0006	0.0078
	Parallel	0.1822	–	0.0302
$n = 20000$	Unparalleled	3.2828	0.0027	0.0075
	Parallel	0.6166	–	0.0293
$n = 500000$	Unparalleled	–	0.1267	0.0508
	Parallel	–	–	0.0374
$n = 1000000$	Unparalleled	–	0.2602	0.1530
	Parallel	–	–	0.0574

Note: All running time in seconds. Parallel computation is conducted over 6 cores. $B = 1000$ when $n \leq 20000$, $B = 3000$ when $n = 500000$, and $B = 5000$ when $n = 1000000$.

ular, for each algorithm, I keep updating 100 times and report the average running time of each single update. For kernel-based updates (KBGD and KMBGD), I consider two computation strategies: unparalleled and parallel computation. When using parallel computation, kernel estimators are simultaneously calculated over 6 cores. I consider six sample sizes: $n = 2500, 5000, 10000, 20000, 500000,$ and 1000000. For SBGD estimation, the sieve functions follow those used in Khan et al. (2023). The order of sieves is chosen as $q = 9$ when $n = 2500$ and 5000, $q = 11$ when $n = 10000$ and 20000, and $q = 31$ when $n = 500000$ and 1000000. The subsample size B is chosen as $B = 1000$ when $n \leq 20000$, $B = 3000$ for $n = 500000$, and $B = 5000$ for $n = 1000000$. The simulation results are reported in Table 2.2.

It can be seen that without parallel computation, the updating time of full-sample-based KBGD algorithm increases roughly at rate n^2, which is in linear with the previous discussion. In particular, when sample size is 2500, each single update requires 0.0475 seconds, which amounts to 21 updates within one second. However, such updating time increases to 0.2 seconds when sample size is 5000, which amounts to only 5 updates each second. When the sample size is 20000, without parallel computation, each single update of KBGD requires more than 3 seconds, indicating that 1000 updates may cost around 1 hour of computational time. For extremely large sample sizes $n = 500000$ or 1000000, KBGD is practically infeasible, so the computational time is not reported. It can also be seen that parallel computation may significantly decrease the updating time when n is large ($n = 10000, 20000$), but the updating time is still too long to be practically feasible.

I then look at the updating speed of SBGD and KMBGD. Apparently, when sample size is small

Table 2.3: Comparing KMBGD and SBGD Estimators

Distribution	Sample Size	Method	RMSE	Running Time	
$u \sim$ Cauchy	$n = 500000$	SBGD	0.0620	0.8417	3.2841
		KMBGD	0.0628	0.4719	0.1042
	$n = 1000000$	SBGD	0.0398	1.7304	13.921
		KMBGD	0.0407	0.5002	0.0968
$u \sim t(4)$	$n = 500000$	SBGD	0.0390	0.8219	3.3434
		KMBGD	0.0390	0.3954	0.1045
	$n = 1000000$	SBGD	0.0273	1.6701	13.893
		KMBGD	0.0276	0.4158	0.4059
$u \sim \chi^2(3)$	$n = 500000$	SBGD	0.0475	0.7016	3.3534
		KMBGD	0.0475	0.4098	0.1047
	$n = 1000000$	SBGD	0.0319	1.4244	14.196
		KMBGD	0.0330	0.3703	0.3515
$u \sim \mathcal{N}(0,1)$	$n = 500000$	SBGD	0.0341	0.8261	3.3310
		KMBGD	0.0341	0.3930	0.1056
	$n = 1000000$	SBGD	0.0216	1.6498	14.134
		KMBGD	0.0218	0.3500	0.3542

NOTE: All running time in hours.

or modest, SBGD exhibits excellent performance: when sample size is 2500, 5000, and 10000, each single update of SBGD requires only 0.0003, 0.0004, and 0.0006 seconds, which amounts to 3300, 2500, and 1600 updates within one second. Even when sample size is 20000, each update of SBGD requires only 0.0027 seconds, so 370 updates can be conducted within one second. This suggests that SBGD significantly outperforms KMBGD when the sample size n is small or modest. However, when the sample size n is extremely large, KMBGD starts dominating SBGD. In particular, when $n = 500000$ and 1000000, the updating speed of KMBGD (with parallel computation) is roughly 4 and 5 times faster than that of SBGD.

Of course, the reduction of computational time of each single update of KMBGD compared with that of SBGD may come at the cost of longer total running time or large estimation error. To study whether it is the case, I then compare the total running time of SBGD and KMBGD. I also consider four setups of random error distributions as I did in subsection 2.4.1. I consider two extreme sample sizes: $n = 500000$ and $n = 1000000$. The subsample size $B = 3000$ when $n = 500000$ and $B = 5000$ when $n = 1000000$. The stopping rule for SBGD is $\max_{1 \leq j \leq 9} |\beta_{j,k+1} - \beta_{j,k}| < 10^{-6}$ and that for KMBGD is the same as before. For both updates, the initial guess is located at Logit estimator, and the maximum number of updates is 20000. For inference, I choose subsample size $B = 3000$ when $n = 500000$ and $B = 6000$ when $n = 1000000$. The number of subsamples is chosen as 200. Finally, I note here that for both estimation and inference, unparalleled computation is considered.

Table 2.4: Comparing True and Estimated Variance

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
$u_i \sim$ Cauchy										
$n = 500000$	True Std	0.0173	0.0102	0.0099	0.0193	0.0442	0.0079	0.0105	0.0159	0.0411
	Est Std	0.0173	0.0099	0.0097	0.0203	0.0444	0.0082	0.0107	0.0177	0.0409
$n = 1000000$	True Std	0.0114	0.0064	0.0069	0.0142	0.0260	0.0057	0.0083	0.0115	0.0264
	Est Std	0.0123	0.0070	0.0068	0.0143	0.0313	0.0058	0.0075	0.0124	0.0287
$u_i \sim t(4)$										
$n = 500000$	True Std	0.0118	0.0059	0.0063	0.0126	0.0280	0.0052	0.0074	0.0113	0.0261
	Est Std	0.0110	0.0062	0.0060	0.0124	0.0275	0.0053	0.0068	0.0111	0.0257
$n = 1000000$	True Std	0.0071	0.0045	0.0040	0.0084	0.0196	0.0041	0.0047	0.0077	0.0180
	Est Std	0.0078	0.0044	0.0043	0.0088	0.0194	0.0037	0.0048	0.0079	0.0182
$u_i \sim \chi^2(3)$										
$n = 500000$	True Std	0.0120	0.0074	0.0066	0.0149	0.0316	0.0067	0.0093	0.0137	0.0321
	Est Std	0.0135	0.0076	0.0071	0.0148	0.0332	0.0068	0.0089	0.0143	0.0325
$n = 1000000$	True Std	0.0092	0.0045	0.0047	0.0107	0.0226	0.0049	0.0061	0.0096	0.0214
	Est Std	0.0096	0.0053	0.0051	0.0105	0.0235	0.0048	0.0063	0.0101	0.0230
$u_i \sim \mathcal{N}(0,1)$										
$n = 500000$	True Std	0.0099	0.0053	0.0049	0.0113	0.0246	0.0048	0.0059	0.0098	0.0225
	Est Std	0.0096	0.0054	0.0053	0.0109	0.0240	0.0046	0.0060	0.0097	0.0225
$n = 1000000$	True Std	0.0068	0.0038	0.0035	0.0072	0.0146	0.0036	0.0040	0.0061	0.0139
	Est Std	0.0068	0.0038	0.0037	0.0077	0.0170	0.0033	0.0042	0.0069	0.0159

I report the RMSE and running time of both estimation and inference in Table 2.3. As can be seen from the table, for all combinations of error distributions and sample sizes, the RMSE of SBGD and KMBGD are almost identical, indicating that updates based on subsamples do not result in loss of estimation accuracy. When looking at the running time, it's impressive to see that, the estimation time of KMBGD is substantially shorter compared with that of SBGD. When $n = 500000$, KMBGD decreases the running time by roughly half, while when n increases to 1000000, the reduction of estimation time is more significant: running time of KMBGD is only around one forth of that of SBGD. It is also interesting to see that, when the sample size increases and I use a larger subsample size, the running time of KMBGD even slightly decreases. This implies that although using a larger subsample size may make updating speed slightly slower, it makes convergence faster because the amount of noises in the update is decreased.

I finally look at the computational burden of inference based on different methods. As can be seen from Table 2.3, the operational time of variance calculation of SBGD is over 3.2 hours without parallel computation when $n = 500000$, and it rises to around 14 hours when $n = 1000000$. This implies that even SBGD may have adequate computational efficiency in terms of estimation, it may still cost a large amount of time to conduct inference. When turning to the subsample-based infernece

under KMBGD, it can be clearly seen that variance estimation only requires around 0.1 hours (10 min) when $n = 500000$ and 0.4 hours (40 min) when $n = 1000000$, which significantly improves the speed of inference. I also report in Table 2.4 the true standard deviation and subsample-based estimator of the standard deviation of each estimator, which are close to each other. This implies that subsample-based inference improves the speed while does not suffer from much accuracy loss.

2.5 Real Data Analysis

2.5.1 `Run_or_walk_information`

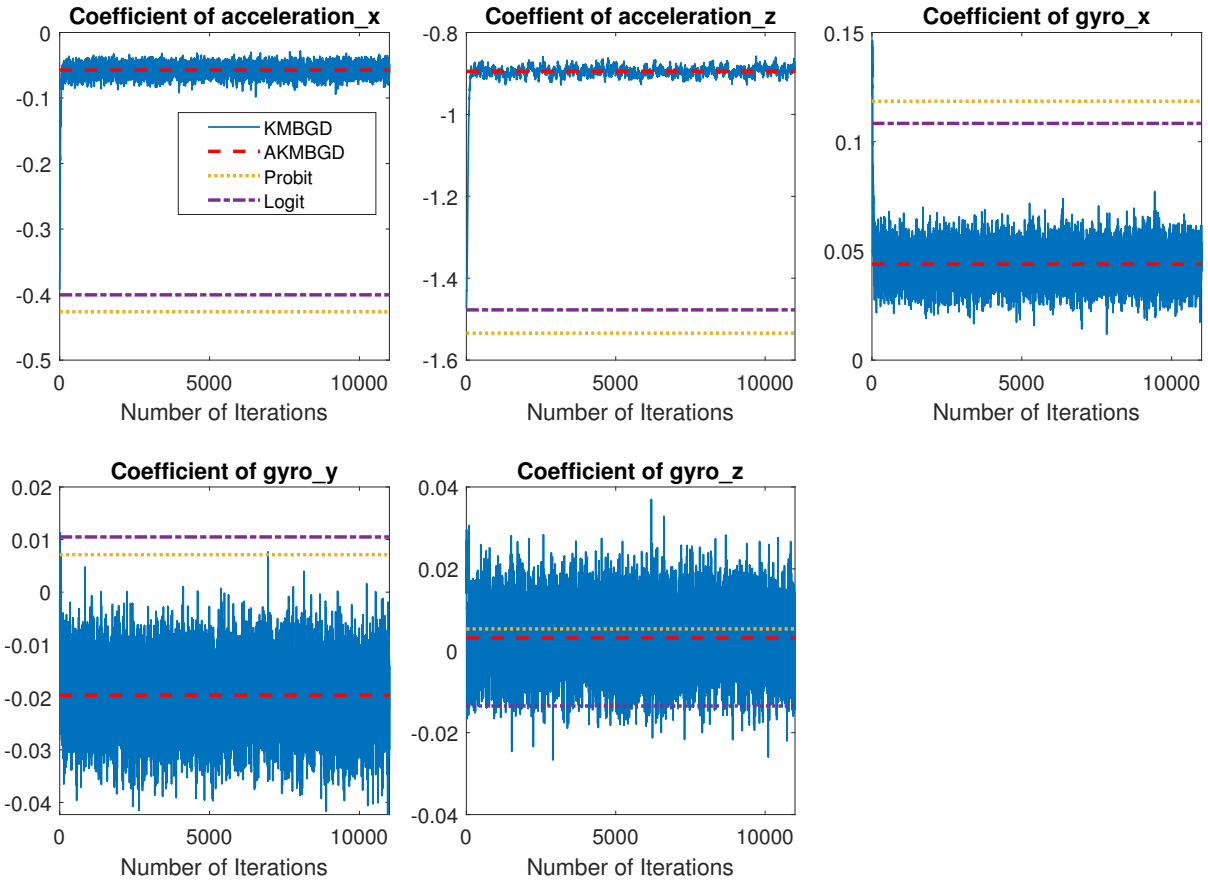
This section applies the KMBGD algorithm to data set `Run_or_walk_information` from OpenML[8]. The data set contains 88,588 observations, each of which has 6 features. So the data set is medium-sized. The binary response y is provided in the data set. We model y and the set of features as a semiparametric binary choice model as in (2.6) and use KMBGD estimation procedure to estimate the model.

When conducing the estimation, we standardize all the covaraites. For the setup of iteration, we choose $\delta_k = 1$, $B = 3000$, $k^* = 10000$, $gap = 1000$, tolerance $\varrho = 0.005$, and the maximum number of iterations as 50000. We normalize the coefficient of `acceleration_y` to be 1 because preliminary Probit and Logit regression indicate that its coefficient is strictly positive. Whenever we construct kernel estimators, we use eighth-order Epanechnikov kernel function given by $K(u) = 16.15(1 - u^2)(0.1667 - 1.5u^2 + 3.3u^4 - 2.043u^6) \cdot \mathbb{1}(|u| \leq 1)$, and the bandwidth is chosen as $h_n = c_k \cdot n^{-1/13}$, where c_k is the standard deviation of the index in the k-th round of iteration. We use Logit estimator as the starting point. Finally, in each update, parallel computation over 6 cores is performed when calculating subsample-based kernel estimators over different data points.

The estimation procedure takes 7.78min in total, where around 11,000 rounds of iterations are conducted. We plot the estimated coefficients based on Probit, Logit, and KMBGD in Figure 2.1. We can see that the KMBGD estimators converge very quickly to be fluctuating closely around the AKMBGD estimatros. Moreover, the estimated coefficients based on KMBGD and AKMBGD differ significantly from those using Probit or Logit, which suggests potential model misspecification under parametric setup.

[8] https://www.openml.org/. Data ID: 40922.

Figure 2.1: Estimated Coefficients of Walk_or_run_information

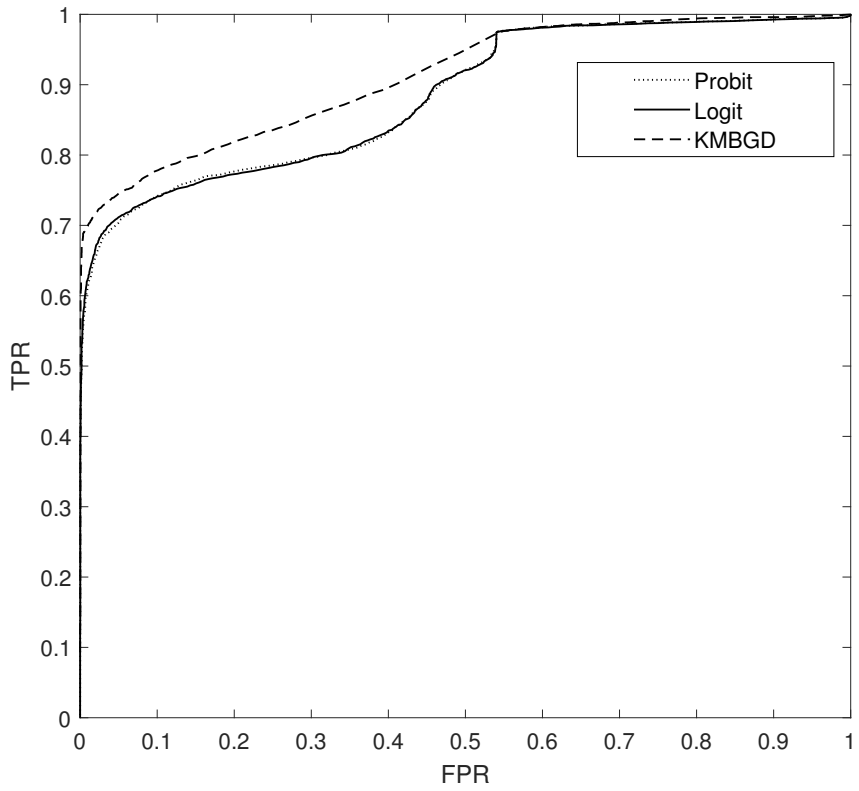


The sizable difference between parametric estimation results and KMBGD estimators suggests potential gain of the use of semiparametric estimation. Observing this, we use different estimation methods to predict the outcome of the binary response variable. In particular, we randomly split the data set into a training set and a testing set, where the latter contains 10,000 observations. Then we use different methods including Probit, Logit, and KMBGD to estimate the training set, and use the estimation results to predict the outcome of the observations in the testing set[9]. We plot the ROC curves of different methods in Figure 2.2. We can see that the ROC curve of KMBGD almost always lies above those of Probit and Logit, indicating better predicting performance. More precisely, the AUC[10] of Probit, Logit and KMBGD are given by 0.8841, 0.8841 and 0.9112, respectively. This implies that KMBGD significantly outperforms the parametric methods in terms of prediction accuracy.

[9]When estimating the conditional probability using AKMBGD estimators, we use second order Epanechnikov kernel function with bandwidth $h_n = c_k \cdot n^{-1/5}$. This also applies to subsection 2.5.2 and subsection 2.5.3.

[10]Area under the ROC curve. The large AUC is, the better prediction accuracy it indicates.

Figure 2.2: ROC of Probit, Logit and KMBGD for Walk_or_run_information



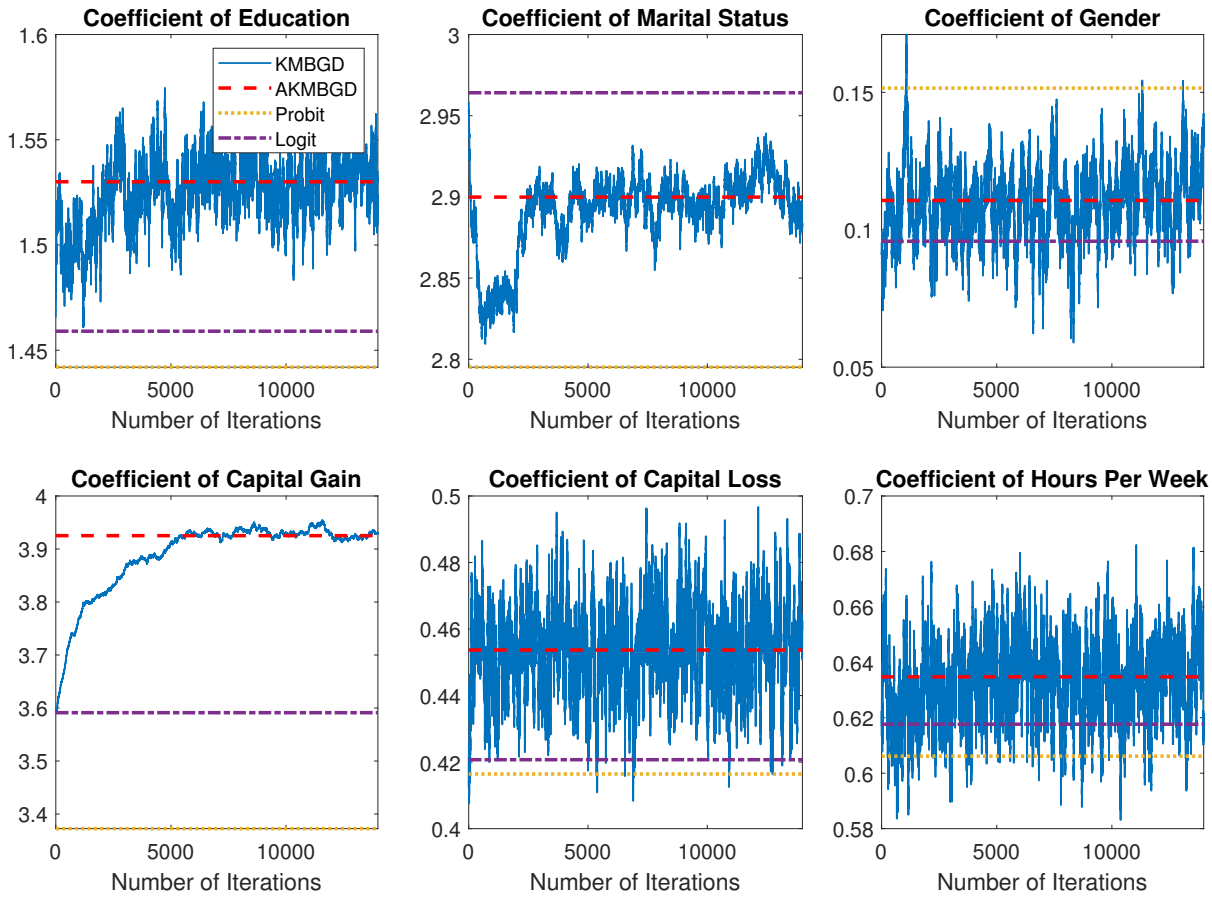
2.5.2 `simulated_adult`

This section applies our method to data set `simulated_adult` from OpenML[11]. The original data set contains 5.1 million observations, each of which has 14 features. The binary response is constructed as $y = 1$ if the class is ">50K" and $y = 0$ otherwise. We model y and the set of features as a semiparametric binary choice model as in (2.6) and use KMBGD estimation procedure to estimate the model. Before we estimate the model, we perform the following data clearing. We leave out observations whose native-country is not United-States, workclass is Without-pay, occupation is Armed-Forces, or race is not White. This leaves us with a data set of 4,734,097 observations, which is an extremely large data set. We generate 5 dummies for workcalss, 12 dummies for occupation, one dummy for marital status, and one dummy for gender[12]. After constructing all the dummies

[11]Data ID: 45689.

[12]We provide more details of the construction of the dummy variables. For workcalss, after dropping Without-pay, we are left with 6 types of workcalss, then we leave out the last type of workcalss for identification. For occupation, after dropping Armed-Forces, we are left with 13 types of occupations, then we leave out the last type of occupation. For marital status, we generate a dummy which is 1 if the marital status is Married-AF-spouse, Married-civ-spouse, or Married-spouse-absent, and is 0 otherwise. Finally, for gender, we generate a dummy that equals 1 if the gender is

Figure 2.3: Partial Estimated Coefficients for `Simulated_adult`



variables, we have 25 regressors in the model, which include age, workclass dummies, fnlwgt, eudcational years, marital status dummy, occupation dummies, gender dummy, capital-gain, capital loss, and hours-per-week. When conducting estimation, we standardize all the covaraites including the dummy variables as we did in the previous section.

For the setup of iteration, we normalize the coefficient of age to be 1 because preliminary Probit and Logit estimation suggest its coefficient is strictly positive. Since we are now working with an extremely large data set, we now choose $B = 5000$. All other setups are the same with those in the previous section.

The estimation procedure takes 16.64min in total. We plot partial estimation results in Figure 2.3. We can see that when we have more covariates, the convergence of the estimated parameters takes more rounds of iterations compared with that in subsection 2.5.1. We can also see that for some

male, and 0 otherwise.

covariates such as Capital Gain, AKMBGD estimators deviate from the Probit or Logit estimators, indicating potential model misspecifications. We finally randomly divide the data set into training and testing sets, where the latter contains 400,000 observations, and compare the prediction accuracy based on different methods. Similar to subsection 2.5.2, we find that the predicting results based on KMBGD are similar to those based on Probit or Logit. The ROC curves of Probit, Logit, and KMBGD almost coincide with each other, with AUC being 0.9227, 0.9226, and 0.9227, respectively.

2.5.3 Revisiting Helpman et al. (2008)

In this section, we will illustrate the empirical applicability of the KMBGD algorithm by revisiting the data set used in Helpman et al. (2008). In their paper, Helpman et al. (2008) consider estimating the following model,

$$\Pr\left(T_{ij}=1|\,\text{observed variables}\right)=G\left(\gamma_0^\star+\xi_j^\star+\zeta_i^\star+\gamma^\star d_{ij}+\kappa^{\star\mathrm{T}}\phi_{ij}\right),$$

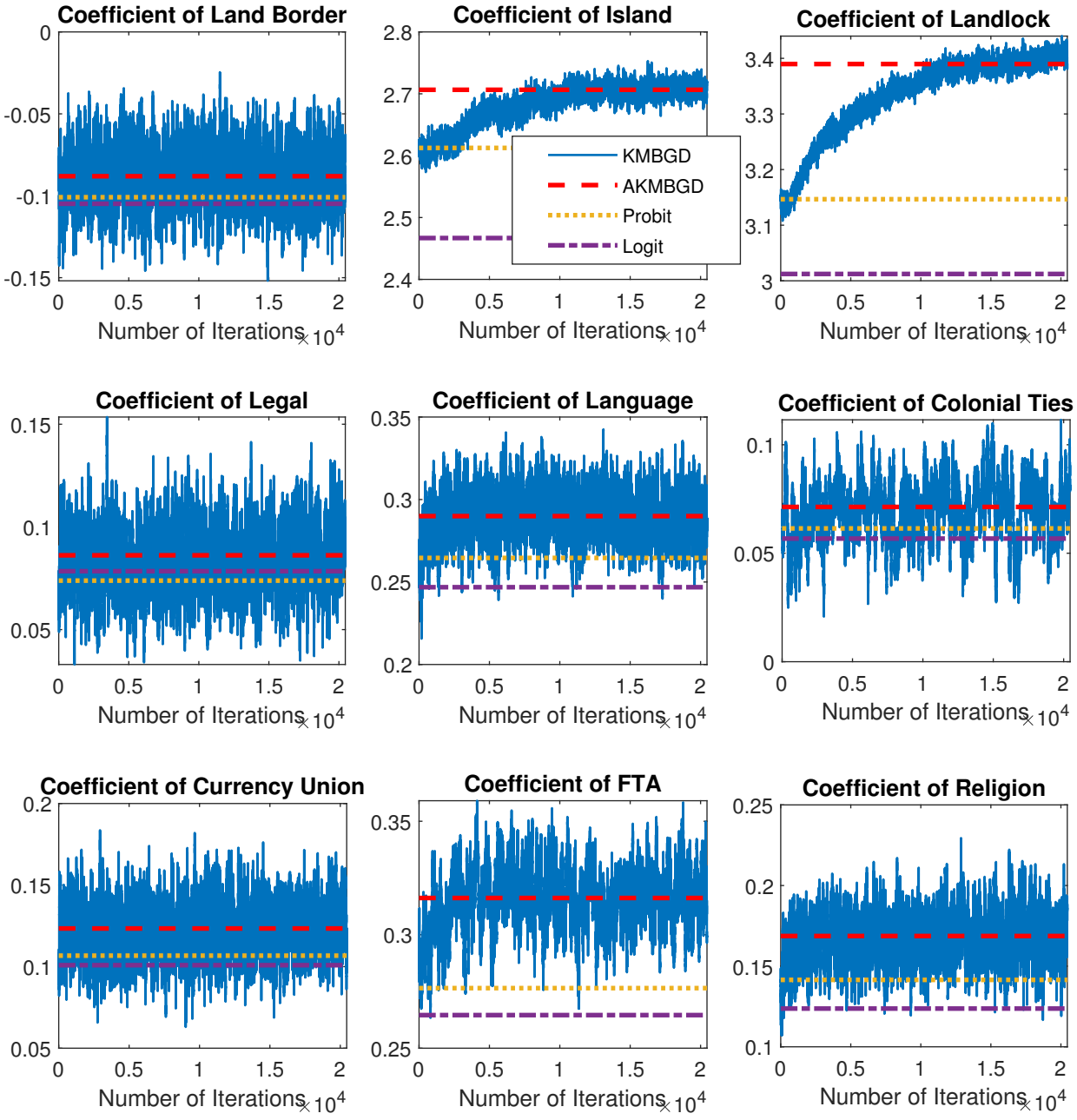
where T_{ij} is an indicator of whether country j exports to country i, $\xi_j^\star$ is the exporter fixed effect of the j-th country, $\zeta_i^\star$ is the importer fixed effect of the i-th country, d_{ij} is the natural logarithm of the geographic distance between countries i and j, and ϕ_{ij} is a vector of covariates that describe the variable country-pair fixed trade cost. The full sample contains a total of 248,060 observations and 336 covariates, which features both large n and p. The covariates contain 10 key variables including Distance, Land Border, Island, Landlock, Legal, Language, Colonial Ties, Currency Union, FTA, and Religion, and 158 exporter fixed effects, 158 importer fixed effects, and 10 year fixed effects.

When estimating the model based on the full sample, Helpman et al. (2008) consider a parametric Probit setup, where G is specified to be the CDF of standard normal distribution. In this section, we reestimate the model without assuming the functional form of G by applying the KMBGD algorithm.

When estimating the model, we standardize all the covaraites including the dummies as we did before. We also leave out as few fixed effects as possible to ensure that the covariate matrix is nonsingular. We choose to normalize the coefficient of negative distance to be 1 since economic theories indicate that a larger geographic distance is generally associated with higher trading costs and such covarite has negative impacts on the conditional probability of the presence of trades between two countries[13] (Helpman et al., 2008). Finally, all the setups of the iteration are the same

[13]When we apply Logit or Probit to estimate the model, the estimated coefficient of Distance is significantly negative.

Figure 2.4: Partial Estimated Coefficients for Data in Helpman et al. (2008)



as those in subsection 2.5.1 except that the initial guess of the parameter is fixed at the Probit estimator.

The estimation procedure takes 25.04min in total. We plot partial estimation results in Figure 2.4. Obviously, since the number of covariates considered in the example is large, the convergence of the estimated parameter is slower compared with that in the previous examples, taking over 20,000 rounds of iterations. We can also see that for some covariates such as Island or Landlock, AKMBGD estimators deviate from the Probit or Logit estimators, indicating potential model misspecifications. We finally randomly divide the data set into training and testing sets, where the latter contains 24806 observations, and compare the prediction accuracy based on different methods. Similar to subsection 2.5.2, we find that the predicting results based on KMBGD are similar to those based on Probit or Logit. The ROC curves of Probit, Logit, and KMBGD almost coincide with each other, with AUC being 0.9388, 0.9391, and 0.9389, respectively.

2.6 Concluding Remarks

This paper investigates semiparametric estimation of monotone index models in a large-n environment, where the number of observations is extremely large. We propose a novel subsample- and iteration-based estimation procedure. Essentially, starting from an initial guess of the parameter, in each round of iteration a subsample is randomly drawn and then used to update the parameter based on the gradient of some well-chosen loss function, where the unknown nonparametric component is replaced with its subsample-based kernel estimator. The proposed algorithm essentially generalizes the idea of mini-batch-based algorithms to the semiparametric setup. Compared with the KBGD algorithm proposed in KLTY, the computational speed of the new estimator substantially improves, so can be easily applied when the sample size n is extremely large. We also show that further averaging across the estimators produced during iterations yields a $1/\sqrt{n}$ consistent and asymptotically normally distributed estimator.

Some issues in this paper remain to be addressed in the future studies. For example, similar to Ichimura (1993), we show that a particular sequence of bandwidth satisfying some order conditions guarantees all the theorems. However, in the theorem the bandwidth is assumed to be unchanged across iterations. Obviously, as the updates proceed, the magnitude of the index value also changes, so a bandwidth adjusted to such change in index value in each round of iteration may lead to a

better kernel estimator and improve the updating results. Similarly, other tuning parameters such as the learning rate δ and subsample size B are all assumed to be given, while their optimal choices remain to be studied.

Another potential future research direction is to generalize the noval subsample-based updating techinque to the full-sample-based SBGD algorithm proposed in KLTY. Different from the kernel-based learning approach, the SBGD algorithm relies on the full sample to update the sieve coefficient in each iteration. So it is still unclear whether using subsamples to perform the update will also yield $1/\sqrt{n}$-consistent estimator. However, since the SBGD algorithm runs significantly faster than the KBGD algorithm, developing subsample-based SBGD algorithm may further improve the computational speed, which deserves further study.

2.7 Appendix

Lemma 2.3. *Suppose that Assumption 2.1–Assumption 2.5 hold with $D \geq 4$. Suppose moreover that $\delta_k = \delta < \min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right), 1/\left(4p^2\left\|G'\right\|_{\infty}\right)\right\}$, $\phi < \delta\underline{\lambda}_{\Lambda}/\left(16p^2\left\|G'\right\|_{\infty}\zeta\right)$, h_n is chosen such that $h_n n^{1/2D} \to 0$ and $h_n n^{1/6}/\log^{1/3}(n) \to \infty$. If $\boldsymbol{\beta}_k$ is updated under (2.4) and (2.10) with $\mathfrak{I}_{B,k} = 1, \cdots, n$, then*

(i) There exists some positive integer k_{KBGD} such that

$$\sup_{k \geq k_{KBGD}} \|\Delta\boldsymbol{\beta}_k\| = O_p\left(n^{-1/2}\right);$$

(ii) Define $\boldsymbol{\xi}_n^{\phi} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{G}\left(z_i^{\star}|\,\boldsymbol{\beta}^{\star}\right) - y_i)\mathbf{X}_i^{\phi}$, where $z_i^{\star} = z\left(\mathbf{X}_{e,i}, \boldsymbol{\beta}^{\star}\right)$. There holds

$$\Delta\boldsymbol{\beta}_{k+1} = \left(I_p - \delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right)\Delta\boldsymbol{\beta}_k - \delta\boldsymbol{\xi}_n^{\phi} + \delta\widetilde{\Omega}_k^{\phi},$$

where $\sup_{k \geq k_{KBGD}}\|\widetilde{\Omega}_k^{\phi}\| = o_p\left(n^{-1/2}\right)$. Define $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_k$ for any k such that $k - k_{KBGD} \to \infty$. There holds $\Delta\widehat{\boldsymbol{\beta}} = -\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\boldsymbol{\xi}_n^{\phi} + o_p(n^{-1/2})$, and

$$\sqrt{n}\Delta\widehat{\boldsymbol{\beta}} \to_d N\left(0, \Sigma_{\boldsymbol{\beta}}^{\phi}\right),$$

where $\Sigma_{\boldsymbol{\beta}}^{\phi}=\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\Sigma_{\boldsymbol{\xi}}^{\phi}\left(\Lambda_{\phi}^{-1}\left(\boldsymbol{\beta}^{\star}\right)\right)^{\mathrm{T}}$ *and*

$$\Sigma_{\boldsymbol{\xi}}^{\phi}=\mathbb{E}\left[\left(1-G\left(z_{i}^{\star}\right)\right)G\left(z_{i}^{\star}\right)\left(\mathbf{X}_{i}^{\phi}-\mathbb{E}\left(\mathbf{X}_{i}^{\phi}\middle|\, z_{i}^{\star}\right)\right)\left(\mathbf{X}_{i}^{\phi}-\mathbb{E}\left(\mathbf{X}_{i}^{\phi}\middle|\, z_{i}^{\star}\right)\right)^{\mathrm{T}}\right].$$

Proof of Lemma 2.3. See Khan et al. (2023). □

Proof of Lemma 2.1

Proof. We start with the proof of the first result. Define $\psi(n,h_n,D)=\sqrt{\log(n)/nh_n}+h_n^D$. Khan et al. (2023) show that

$$\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|\widehat{G}\left(z|\,\boldsymbol{\beta}\right)-\mathbb{E}\left(y|\,X_0+\mathbf{X}^T\boldsymbol{\beta}=z\right)\right|=O_p\left(\psi\left(n,h_n,D\right)\right).$$

Define event

$$e_{1,n}=\left\{\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|\widehat{G}\left(z|\,\boldsymbol{\beta}\right)\right|\leq 2\right\},$$

then $P\left(e_{1,n}\right)\to 1$ since $\psi\left(n,h_n,D\right)\to 0$ according to the choice of h_n. Over event $e_{1,n}$, we have that

$$\mathbb{E}_k^{*}\left\|\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}\left(\widehat{G}\left(X_0+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k|\,\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi}-\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(X_0+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k|\,\boldsymbol{\beta}_k\right)-y_i\right)\mathbf{X}_i^{\phi}\right\|\leq\frac{C}{B}.$$

Now we prove the second result. Recall that $A_{n,y}\left(z,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)y_i$, $A_{n,1}\left(z,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $A_{n,y}\left(z,\boldsymbol{\beta}|\,\mathfrak{I}_{B,k}\right)=\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)y_i$, and $A_{n,1}\left(z,\boldsymbol{\beta}|\,\mathfrak{I}_{B,k}\right)=\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}}K_{h_n}\left(z-X_{0,i}-\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$. According to Khan et al. (2023),

$$\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|A_{n,1}\left(z,\boldsymbol{\beta}\right)-f_Z\left(z|\,\boldsymbol{\beta}\right)\right|=O_p\left(\psi\left(n,h_n,D\right)\right).$$

Note that $\inf_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}f_Z\left(z|\,\boldsymbol{\beta}\right)\geq 3\underline{c}_f$ and $\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}f_Z\left(z|\,\boldsymbol{\beta}\right)\leq\overline{c}_f$, where $\overline{c}_f$ is some sufficiently large positive constant, define event

$$e_{2,n}=\left\{2\underline{c}_f\leq\inf_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}A_{n,1}\left(z,\boldsymbol{\beta}\right)\leq\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}A_{n,1}\left(z,\boldsymbol{\beta}\right)\leq 2\overline{c}_f\right\}.$$

Since $\psi\left(n,h_n,D\right)\to 0$, we have that $P\left(e_{2,n}\right)\to 1$. Moreover, $P\left(e_{1,n}\cap e_{2,n}\right)\to 1$ and over $e_{1,n}\cap e_{2,n}$,

we have that

$$\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|A_{n,y}\left(z,\boldsymbol{\beta}\right)\right|\leq\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|A_{n,1}\left(z,\boldsymbol{\beta}\right)\right|\cdot\sup_{z\in\mathcal{Z}^{\phi},\boldsymbol{\beta}\in\mathcal{B}}\left|\widehat{G}\left(z|\,\boldsymbol{\beta}\right)\right|\leq4\overline{c}_{f}.$$

Define

$$e_{3,n,k}^{\epsilon}=\left\{\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z,\boldsymbol{\beta}_{k}\right)\right|<\epsilon\right\}$$

and

$$e_{4,n,k}^{\epsilon}=\left\{\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)\right|<\epsilon\right\}.$$

For $\epsilon=\epsilon\left(\zeta\right)=2\underline{c}_{f}/\zeta$ with $\zeta>2$, we have that over $e_{1,n}\cap e_{2,n}\cap e_{3,n,t}^{\epsilon}\cap e_{4,n,t}^{\epsilon}$, there holds

$$\begin{aligned}&\sup_{z\in\mathcal{Z}^{\phi}}\left|\frac{A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)}{A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)}-\frac{A_{n,y}\left(z,\boldsymbol{\beta}_{k}\right)}{A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)}\right|\\&\leq\sup_{z\in\mathcal{Z}^{\phi}}\left|\frac{A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z,\boldsymbol{\beta}_{k}\right)}{A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)}\right|+\sup_{z\in\mathcal{Z}^{\phi}}\left|\frac{A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)\left(A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)\right)}{A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)}\right|\\&\leq\frac{1}{2\underline{c}_{f}}\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z,\boldsymbol{\beta}_{k}\right)\right|+\frac{4\overline{c}_{f}+2\underline{c}_{f}/\zeta}{\left(2\underline{c}_{f}\right)\left(2\underline{c}_{f}-2\underline{c}_{f}/\zeta\right)}\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)\right|\\&\leq c_{1}\left(\zeta\right)\epsilon,\end{aligned}$$

where

$$c_{1}\left(\zeta\right)=\frac{1}{2\underline{c}_{f}}+\frac{4\overline{c}_{f}\zeta+2\underline{c}_{f}}{4\underline{c}_{f}^{2}\left(\zeta-1\right)}\leq c_{1}^{\infty},$$

and c_{1}^{∞} is a positive constant depending only on $\overline{c}_{f}$ and $\underline{c}_{f}$. Moreover, when $\epsilon=\underline{c}_{f}/\zeta$ is chosen such that $\zeta>2$, there holds $2\underline{c}_{f}/\zeta<\underline{c}_{f}$, so over $e_{1,n}\cap e_{2,n}\cap e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}$, there holds $\inf_{z\in\mathcal{Z}^{\phi}}A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)\geq\underline{c}_{f}$, and $\widehat{G}\left(z|\,\boldsymbol{\beta}_{k},\mathfrak{I}_{B,k},\underline{c}_{f}\right)=A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)/A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)$.

Since $\left|K_{h_{n}}\left(z-X_{0,i}-\mathbf{X}_{i}^{\mathrm{T}}\boldsymbol{\beta}_{k}\right)\right|\leq Ch_{n}^{-1}$, we have that for any fixed z and ϵ,

$$\mathbb{P}_{k}^{*}\left(\left|A_{n,1}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_{k}\right)\right|>\epsilon\right)\leq2\exp\left(-CBh_{n}^{2}\epsilon^{2}/2\right),$$

and

$$\mathbb{P}_{k}^{*}\left(\left|A_{n,y}\left(z,\boldsymbol{\beta}_{k}|\,\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z,\boldsymbol{\beta}_{k}\right)\right|>\epsilon\right)\leq2\exp\left(-CBh_{n}^{2}\epsilon^{2}/2\right),$$

Also note that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,1}\left(z,\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_k\right)\right| \\
&\leq \max_{1\leq s\leq S}\left|A_{n,1}\left(z_s,\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z_s,\boldsymbol{\beta}_k\right)\right|+Ch_n^{-2}/S,
\end{aligned}$$

for any positive integer S and a set of well-chosen points $z_1,\cdots,z_S$ in $\mathcal{Z}^{\phi}$, where the positive constant C does not depend on $\boldsymbol{\beta}_k$, the index set $\mathfrak{I}_{B,k}$, S, and the choice of $z_1,\cdots,z_S$. Let S be such that $Ch_n^{-2}/S<\epsilon$, we have that

$$\begin{aligned}
&\mathbb{P}_k^*\left(\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,1}\left(z,\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z,\boldsymbol{\beta}_k\right)\right|>\epsilon\right) \\
&\leq\sum_{s=1}^{S}\mathbb{P}_k^*\left(\left|A_{n,1}\left(z_s,\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,1}\left(z_s,\boldsymbol{\beta}_k\right)\right|>\epsilon-Ch_n^{-2}/S\right) \\
&\leq 2\exp\left(\log S-Bh_n^2\left(\epsilon-Ch_n^{-2}/S\right)^2/2\right). \qquad (2.15)
\end{aligned}$$

Using similar method, we can show that

$$\begin{aligned}
&\mathbb{P}_k^*\left(\sup_{z\in\mathcal{Z}^{\phi}}\left|A_{n,y}\left(z,\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z,\boldsymbol{\beta}_k\right)\right|>\epsilon\right) \\
&\leq 2\exp\left(\log S-Bh_n^2\left(\varepsilon-Ch_n^{-2}/S\right)^2/2\right). \qquad (2.16)
\end{aligned}$$

Now consider $\mathbb{E}_k^*\left\|\pi_{2,n,k}\right\|^2$ when $e_{1,n}\cap e_{2,n}$ occurs. We first have that

$$\begin{aligned}
\mathbb{E}_k^*\left\|\pi_{2,n,k}\right\|^2=&\,\mathbb{E}_k^*\left(\left\|\pi_{2,n,k}\right\|^2\middle|e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)\mathbb{P}_k^*\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right) \\
&+\mathbb{E}_k^*\left(\left\|\pi_{2,n,k}\right\|^2\middle|\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)^C\right)\mathbb{P}_k^*\left(\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)^C\right).
\end{aligned}$$

For $\epsilon<2\underline{c}_f/\zeta$ with $\zeta>2$, we have that

$$\mathbb{E}_k^*\left(\left\|\pi_{2,n,k}\right\|^2\middle|e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)\leq c_1^{\infty 2}\left\|\mathbf{X}^{\phi}\right\|_{\infty}^2\epsilon^2=C\epsilon^2.$$

On the other side, according to (2.15) and (2.16), we have that

$$\begin{aligned}
&\mathbb{E}_k^*\left(\left\|\pi_{2,n,k}\right\|^2\middle|\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)^C\right)\mathbb{P}_k^*\left(\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)^C\right) \\
&\leq Ch_n^{-2}\mathbb{P}_k^*\left(\left(e_{3,n,k}^{\epsilon}\cap e_{4,n,k}^{\epsilon}\right)^C\right)\leq Ch_n^{-2}\exp\left(C\log S-CBh_n^2\left(\epsilon-Ch_n^{-2}/S\right)^2/2\right).
\end{aligned}$$

Together we have that over $e_{1,n} \cap e_{2,n}$, there holds

$$\mathbb{E}_k^* \left\| \pi_{2,n,k} \right\|^2 \leq C \left(\epsilon^2 + h_n^{-2} \exp \left(C \log S - C B h_n^2 \left(\epsilon - C h_n^{-2} / S \right)^2 / 2 \right) \right).$$

If we choose

$$S = 2C \sqrt{\frac{B h_n^{-2}}{\log \left(B h_n^{-2} \right)}}, \ \epsilon = \sqrt{\frac{8 \left(\log \left(h_n^{-2} \right) + \log \left(4 C^2 B h_n^{-2} \right) + \log \left(8 B h_n^2 \right) \right)}{B h_n^2}},$$

we have that $C h_n^{-2} / S \leq \epsilon / 2$ and $\epsilon < 2 \underline{c}_f$ for n sufficiently large, and

$$\mathbb{E}_k^* \left\| \pi_{2,n,k} \right\|^2 \leq C \frac{\log \left(B h_n^{-2} \right)}{B h_n^2}.$$

Since $\sup_{k \geq 1} \mathbb{E}_k^* \left\| \pi_{2,n,k} \right\|^2 \leq C$ implies that $\sup_{k \geq 1} \mathbb{E}^* \left\| \pi_{2,n,k} \right\|^2 \leq C$, we have that

$$\begin{aligned} P \left(\sup_{k \geq 1} \mathbb{E}^* \left\| \pi_{2,n,k} \right\|^2 \leq C \frac{\log \left(B h_n^{-2} \right)}{B h_n^2} \right) &\geq P \left(\sup_{k \geq 1} \mathbb{E}_k^* \left\| \pi_{2,n,k} \right\|^2 \leq C \frac{\log \left(B h_n^{-2} \right)}{B h_n^2} \right) \\ &\geq P \left(e_{1,n} \cap e_{2,n} \right) \to 1. \end{aligned}$$

This proves the result. □

Proof of Theorem 2.1

Proof. Note that

$$\begin{aligned} \left\| \Delta \boldsymbol{\beta}_{k+1} \right\| &\leq \sup_{\boldsymbol{\beta} \in \mathcal{B}} \overline{\sigma} \left(I_p - \delta \Lambda_\phi \left(\boldsymbol{\beta} \right) \right) \left\| \Delta \boldsymbol{\beta}_k \right\| + \delta \left(\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \eta_{1,n} \left(\boldsymbol{\beta} \right) \right\| + \left\| \eta_{2,n} \right\| + \left\| \pi_{1,n,k} \right\| + \left\| \pi_{2,n,k} \right\| \right) \\ &\leq \left(1 - \delta \underline{\lambda}_\Lambda / 16 \right) \left\| \Delta \boldsymbol{\beta}_k \right\| + \delta \left(\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\| \eta_{1,n} \left(\boldsymbol{\beta} \right) \right\| + \left\| \eta_{2,n} \right\| + \left\| \pi_{1,n,k} \right\| + \left\| \pi_{2,n,k} \right\| \right), \end{aligned}$$

where

$$\eta_{1,n} \left(\boldsymbol{\beta} \right) = \frac{1}{n} \sum_{i=1}^n \widehat{G} \left(z \left(\mathbf{X}_{e,i}, \boldsymbol{\beta} \right) | \boldsymbol{\beta} \right) \mathbf{X}_i - \mathbb{E} \left[L \left(z \left(\mathbf{X}_{e,i}, \boldsymbol{\beta} \right), \boldsymbol{\beta} \right) \mathbf{X}_i \right],$$

$$\eta_{2,n} = \left(\frac{1}{n} \sum_{i=1}^n G \left(z_i^\star \right) \mathbf{X}_i - \mathbb{E} \left[G \left(z_i^\star \right) \mathbf{X}_i \right] \right) + \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot \mathbf{X}_i.$$

Using Minkovski inequality, we have that

$$\begin{aligned}\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k+1}\right\|^2\right)^{1/2} &\leq (1-\delta\underline{\lambda}_{\Lambda}/16)\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_k\right\|^2\right)^{1/2}+\delta\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\delta\left\|\eta_{2,n}\right\| \\ &\quad+\delta\left(\mathbb{E}^*\left\|\pi_{1,n,k}\right\|^2\right)^{1/2}+\delta\left(\mathbb{E}^*\left\|\pi_{2,n,k}\right\|^2\right)^{1/2} \\ &\leq (1-\delta\underline{\lambda}_{\Lambda}/16)\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_k\right\|^2\right)^{1/2}+\delta\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\delta\left\|\eta_{2,n}\right\| \\ &\quad+CB^{-1/2}+C\left(\frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right)^{1/2}.\end{aligned}$$

This implies that

$$\begin{aligned}\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k+1}\right\|^2\right)^{1/2} &\leq (1-\delta\underline{\lambda}_{\Lambda}/16)^k\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_1\right\|^2\right)^{1/2} \\ &\quad+C\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\left\|\eta_{2,n}\right\|+\left(\frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right)^{1/2}\right).\end{aligned}$$

Then when $k \geq k_n + 1$, we have that

$$(1-\delta\underline{\lambda}_{\Lambda}/16)^k\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_1\right\|^2\right)^{1/2} \leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\left\|\eta_{2,n}\right\|+\left(\log\left(Bh_n^{-2}\right)/Bh_n^2\right)^{1/2},$$

implying that $\left(\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k+1}\right\|^2\right)^{1/2} = O_p\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\left\|\eta_{2,n}\right\|+\left(\log\left(Bh_n^{-2}\right)/Bh_n^2\right)^{1/2}\right)$. Finally, Khan et al. (2023) show that $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\eta_{1,n}(\boldsymbol{\beta})\right\|+\left\|\eta_{2,n}\right\| = O_p(\psi(n,h_n,D))$. Since $B \leq n$, we have that

$$\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k+1}\right\|^2 = O_p\left(h_n^{2D}+\frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right).$$

□

Proof of Lemma 2.2

Proof. Note that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+1} = & \int_0^1 \left(I_p - \delta \Lambda_\phi\left(\boldsymbol{\beta}^\star + \tau\Delta\boldsymbol{\beta}_k\right)\right) d\tau\Delta\boldsymbol{\beta}_k - \delta\boldsymbol{\xi}_n^\phi \\
& - \delta \int_0^1 \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\phi \frac{\partial \widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} \middle| \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\star+\tau\Delta\boldsymbol{\beta}_k} - \Lambda_\phi\left(\boldsymbol{\beta}^\star + \tau\Delta\boldsymbol{\beta}_k\right)\right) d\tau\Delta\boldsymbol{\beta}_k (i) \\
& - \delta \left(\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^\phi - \frac{1}{n}\sum_{i=1}^n \left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i^\phi \right) (ii) \\
& - \delta \left(\frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k, \mathfrak{I}_{B,k}, \underline{c}_f\right) - \widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right)\right)\mathbf{X}_i^\phi \right) (iii).
\end{aligned}$$

For (i), we have that

$$\begin{aligned}
\sup_{k\geq k_n+1} \mathbb{E}_k^* \left\|(i)\right\| &= O_p\left(\left(h_n^{-2}\sqrt{\frac{\log(n)}{n}} + h_n^D\right)\left(h_n^D + \sqrt{\frac{\log(n)}{Bh_n^2}}\right) + h_n^{2D} + \frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2}\right) \\
&= O_p\left(\sqrt{\frac{\log^2(n)}{nBh_n^6}} + h_n^{D-2}\sqrt{\frac{\log(n)}{n}} + \frac{\log\left(Bh_n^{-2}\right)}{Bh_n^2} + h_n^{2D}\right).
\end{aligned}$$

This implies that given the choice of B and h_n, $\mathbb{E}^*\left\|(i)\right\|$ is $o_p(n^{-1/2})$ uniformly with respect to k.

Now we look at (iii). To further simplify our notations, we denote $A_{n,y}\left(z_{i,k},\boldsymbol{\beta}_k\right) = A_{n,y,i,k}$, $A_{n,1}\left(z_{i,k},\boldsymbol{\beta}_k\right) = A_{n,1,i,k}$, $A_{n,y}\left(z_{i,k},\boldsymbol{\beta}_k\middle|\mathfrak{I}_{B,k}\right) = A_{n,y,i,k}^{\mathfrak{I}}$, $A_{n,1}\left(z_{i,k},\boldsymbol{\beta}_k\middle|\mathfrak{I}_{B,k}\right) = A_{n,1,i,k}^{\mathfrak{I}}$. We have that

$$\begin{aligned}
(iii) = & \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \left(\frac{A_{n,y,i,k}^{\mathfrak{I}}}{A_{n,1,i,k}^{\mathfrak{I}} \wedge \underline{c}_f} - \frac{A_{n,y,i,k}}{A_{n,1,i,k}}\right)\mathbf{X}_i^\phi \\
= & \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{\mathbf{X}_i^\phi}{A_{n,1,i,k}} \cdot \left(A_{n,y,i,k}^{\mathfrak{I}} - A_{n,y,i,k}\right) (iv) \\
& - \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{A_{n,y,i,k}\mathbf{X}_i^\phi}{A_{n,1,i,k}^2}\left(A_{n,1,i,k}^{\mathfrak{I}} \wedge \underline{c}_f - A_{n,1,i,k}^{\mathfrak{I}}\right) (v) - \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{A_{n,y,i,k}\mathbf{X}_i^\phi}{A_{n,1,i,k}^2}\left(A_{n,1,i,k}^{\mathfrak{I}} - A_{n,1,i,k}\right) (vi) \\
& - \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{\mathbf{X}_i^\phi}{\widetilde{A}_{n,1,i,k}^2}\left(A_{n,y,i,k}^{\mathfrak{I}} - A_{n,y,i,k}\right)\left(A_{n,1,i,k}^{\mathfrak{I}} \wedge \underline{c}_f - A_{n,1,i,k}^{\mathfrak{I}}\right) (vii) \\
& - \frac{1}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{\mathbf{X}_i^\phi}{\widetilde{A}_{n,1,i,k}^2}\left(A_{n,y,i,k}^{\mathfrak{I}} - A_{n,y,i,k}\right)\left(A_{n,y,i,k}^{\mathfrak{I}} - A_{n,y,i,k}\right) (viii) \\
& + \frac{2}{B}\sum_{i\in\mathfrak{I}_{B,k}} \frac{A_{n,1,i,k}\mathbf{X}_i^\phi}{\widetilde{\widetilde{A}}_{n,1,i,k}^3}\left(A_{n,y,i,k}^{\mathfrak{I}} - A_{n,y,i,k}\right)^2 (ix),
\end{aligned}$$

where $\widetilde{A}^2_{n,1,i,k}$ and $\widetilde{\widetilde{A}}^3_{n,1,i,k}$ both lie between $A^{\mathfrak{J}}_{n,1,i,k} \wedge \underline{c}_f$ and $A_{n,1,i,k}$. Define $mathbbE^*_k\{|j\}$ as the conditional expectation with respect to $\mathbb{P}^*_k$ holding the j-th index $i_{k,j}$ fixed. Note that for any $1 \leq j \leq B$ and k,

$$\begin{aligned}
&\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,y,i_{k,j},k} - A_{n,y,i_{k,j},k} \right)^2 \middle| j \right\} \\
&= \mathbb{E}^*_k \left\{ \left(\frac{1}{B} \sum_{b=1}^{B} K_{h_n} \left(z_{i_{k,j},k} - z_{i_{k,b},k} \right) y_{i_b} - \frac{1}{n} \sum_{b=1}^{n} K_{h_n} \left(z_{i_{k,j},k} - z_{b,k} \right) y_{k,b} \right)^2 \middle| j \right\} \\
&\leq C \left\{ \left(\frac{y_{i_{k,j}}}{Bh_n} - \frac{1}{Bn} \sum_{b=1}^{n} K_{h_n} \left(z_{i_{k,j},k} - z_{b,k} \right) y_b \right)^2 + \frac{B-1}{B^2} \frac{1}{n} \sum_{b=1}^{n} K^2_{h_n} \left(z_{i_{k,j},k} - z_{b,k} \right) y_b^2 \right\} \leq \frac{C}{Bh_n^2},
\end{aligned}$$

for some positive constant C that does not depend on k and j. Similarly, we have that for all $1 \leq j \leq B$ and k,

$$\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} - A_{n,1,i_{k,j},k} \right)^2 \middle| j \right\} \leq \frac{C}{Bh_n^2}.$$

So with probability going to 1, for all k

$$\begin{aligned}
\mathbb{E}^*_k \left\| (viii) \right\| &\leq \frac{C}{B} \mathbb{E}^*_k \left(\sum_{i \in \mathfrak{I}_{B,k}} \left| \left(A^{\mathfrak{J}}_{n,y,i,k} - A_{n,y,i,k} \right) \left(A^{\mathfrak{J}}_{n,1,i,k} - A_{n,1,i,k} \right) \right| \right) \\
&\leq \frac{C}{B} \mathbb{E}^*_k \left(\sum_{j=1}^{B} \mathbb{E}^*_k \left(\left| \left(A^{\mathfrak{J}}_{n,y,i_{k,j},k} - A_{n,y,i_{k,j},k} \right) \left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} - A_{n,1,i_{k,j},k} \right) \right| \middle| j \right) \right) \\
&\leq \frac{C}{B} \mathbb{E}^*_k \left(\sum_{j=1}^{B} \sqrt{\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,y,i_{k,j},k} - A_{n,y,i_{k,j},k} \right)^2 \middle| j \right\}} \sqrt{\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} - A_{n,1,i_{k,j},k} \right)^2 \middle| j \right\}} \right) \\
&\leq \frac{C}{B} \mathbb{E}^*_k \left(\sum_{j=1}^{B} \frac{C}{Bh_n^2} \right) \leq \frac{C}{Bh_n^2}.
\end{aligned}$$

Similarly, we have that $\mathbb{E}^*_k \left\| (ix) \right\| \leq C/Bh_n^2$ for all k with probability going to 1. Due to the choice of B and h_n, we have that $\mathbb{E}^* \left\| (viii) \right\|$ and $\mathbb{E}^* \left\| (ix) \right\|$ are both $o_p(n^{-1/2})$ uniformly with respect to k. On the other side, note that

$$\begin{aligned}
\mathbb{E}^*_k \left\| (vii) \right\| &\leq C \mathbb{E}^*_k \left(\frac{1}{B} \sum_{j=1}^{B} \sqrt{\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,y,i_{k,j},k} - A_{n,y,i_{k,j},k} \right)^2 \middle| j \right\}} \sqrt{\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} \wedge \underline{c}_f - A_{n,1,i_{k,j},k} \right)^2 \middle| j \right\}} \right) \\
&\leq C \mathbb{E}^*_k \left(\frac{1}{B} \sum_{j=1}^{B} \left(\frac{C}{\sqrt{Bh_n^2}} \right) \sqrt{\mathbb{E}^*_k \left\{ \left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} \wedge \underline{c}_f - A_{n,1,i_{k,j},k} \right)^2 \middle| j \right\}} \right).
\end{aligned}$$

Note that

$$\mathbb{E}_k^*\left\{\left(A^{\mathfrak{J}}_{n,1,i_{k,j},k}\wedge \underline{c}_f - A_{n,1,i_{k,j},k}\right)^2\middle| j\right\} \leq Ch_n^{-2}\mathbb{P}_k^*\left(A^{\mathfrak{J}}_{n,1,i_j,k} < \underline{c}_f\middle| j\right).$$

Now consider $\mathbb{P}_k^*\left(A^{\mathfrak{J}}_{n,1,i_{k,j},k} < \underline{c}_f\middle| j\right)$. Note that

$$\begin{aligned}
A^{\mathfrak{J}}_{n,1,i_{k,j},k} < \underline{c}_f \Longrightarrow & \frac{1}{B}\sum_{b=1}^{B} K_{h_n}\left(z_{i_{k,j},k} - z_{i_b}\right) y_{i_b} - \frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(z_{i_{k,j},k} - z_{i,k}\right) y_i \\
& < \underline{c}_f - \frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(z_{i_{k,j},k} - z_{i,k}\right) y_i \\
\Longrightarrow & \frac{1}{B}\sum_{b\neq j}^{B} K_{h_n}\left(z_{i_{k,j},k} - z_{i_{k,b}}\right) y_{i_{k,b}} - \frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(z_{i_{k,j},k} - z_{i,k}\right) y_i < -\underline{c}_f - \frac{y_{i_{k,j}}}{Bh_n} \\
\Longrightarrow & \sup_{z\in\mathcal{Z}^\phi}\left|\frac{1}{B}\sum_{b\neq j}^{B} K_{h_n}\left(z_{i_{k,j},k} - z_{i_{k,b}}\right) y_{i_{k,b}} - \frac{B-1}{B}\frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(z_{i_{k,j},k} - z_{i,k}\right) y_i\right| > \underline{c}_f + \frac{C}{Bh_n}.
\end{aligned}$$

This implies that

$$\begin{aligned}
& \mathbb{P}_k^*\left(A^{\mathfrak{J}}_{n,1,i_j,k} < \underline{c}_f\middle| j\right) \\
& \leq \mathbb{P}_k^*\left(\sup_{z\in\mathcal{Z}}\left|\frac{1}{B}\sum_{b\neq j}^{B} K_{h_n}\left(z_{i_{k,j},k} - z_{i_{k,b}}\right) y_{i_{k,b}} - \frac{B-1}{B}\frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(z_{i_{k,j},k} - z_{i,k}\right) y_i\right| > \underline{c}_f + \frac{C}{Bh_n}\middle| j\right) \\
& \leq 2\exp\left(\log S - Bh_n^2\left(\underline{c}_f + \frac{C}{Bh_n} - Ch_n^{-2}/S\right)^2/2\right)
\end{aligned}$$

for any sufficiently large positive integer S. Let $S = Bh_n^{-1}$, we have that for n sufficiently large, we have that

$$\exp\left(\log S - Bh_n^2\left(\underline{c}_f - \frac{C}{Bh_n} + \frac{C}{h_n^2 S}\right)^2/2\right) \leq C\exp\left(C\left(\log\left(Bh_n^{-1}\right) - Bh_n^2\right)\right),$$

implying that

$$\mathbb{E}_k^*\left\{\left(A^{\mathfrak{J}}_{n,1,i_{k,j},k}\wedge \underline{c}_f - A_{n,1,i_{k,j},k}\right)^2\middle| j\right\} \leq Ch_n^{-2}\exp\left(C\left(\log\left(Bh_n^{-1}\right) - Bh_n^2\right)\right).$$

So uniformly with respect to k, there holds

$$\mathbb{E}_k^*\left\|(vii)\right\| \leq \frac{C\exp\left(C\left(\log\left(Bh_n^{-1}\right) - Bh_n^2\right)\right)}{\sqrt{Bh_n^4}}.$$

Similarly, we have that $\mathbb{E}_k^*\left\|(v)\right\| \leq Ch_n^{-1}\exp\left(C\left(\log\left(Bh_n^{-1}\right) - Bh_n^2\right)\right)$ for all k. Given the choice of

B and h_n, we have that $\mathbb{E}^* \|(vii)\|$ and $\mathbb{E}^* \|(v)\|$ are both $o_p(n^{-1/2})$ uniformly with respect to k.

We finally note that uniformly for all k,

$$\mathbb{E}^* \left\| \left(\int_0^1 \Lambda_\phi \left(\boldsymbol{\beta}^\star + \tau \Delta \boldsymbol{\beta}_k \right) d\tau - \Lambda_\phi \left(\boldsymbol{\beta}^\star \right) \right) \Delta \boldsymbol{\beta}_k \right\| \leq C \mathbb{E}^* \left\| \Delta \boldsymbol{\beta}_k \right\|^2 = O_p \left(h_n^{2D} + \frac{\log \left(B h_n^{-2} \right)}{B h_n^2} \right).$$

This finishes the proof. □

Proof of Theorem 2.2

Proof. Define

$$\boldsymbol{\Xi}_{1,k}^{\phi} = \frac{1}{B} \sum_{i \in \mathfrak{I}_{B,k}} \left(\widehat{G} \left(z_{i,k} | \, \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_i^{\phi} - \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{G} \left(z_{i,k} | \, \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_i^{\phi},$$

$$\boldsymbol{\Xi}_{2,k}^{\phi} = \frac{1}{B} \sum_{i \in \mathfrak{I}_{B,k}} \frac{\mathbf{X}_i^{\phi}}{A_{n,1} \left(z_{i,k}, \boldsymbol{\beta}_k \right)} \left(A_{n,y} \left(z_{i,k}, \boldsymbol{\beta}_k | \, \mathfrak{I}_{B,k} \right) - A_{n,y} \left(z_{i,k}, \boldsymbol{\beta}_k \right) \right),$$

and

$$\boldsymbol{\Xi}_{3,k}^{\phi} = \frac{1}{B} \sum_{i \in \mathfrak{I}_{B,k}} \frac{A_{n,y} \left(z_{i,k}, \boldsymbol{\beta}_k \right) \mathbf{X}_i^{\phi}}{A_{n,1}^2 \left(z_{i,k}, \boldsymbol{\beta}_k \right)} \left(A_{n,1} \left(z_{i,k}, \boldsymbol{\beta}_k | \, \mathfrak{I}_{B,k} \right) - A_{n,1} \left(z_{i,k}, \boldsymbol{\beta}_k \right) \right).$$

We obviously have that $\sup_k \mathbb{E}_k^* \left\| \boldsymbol{\Xi}_{1,k}^{\phi} \right\|^2 \leq C/B$, so $\sup_k \mathbb{E}^* \left\| \boldsymbol{\Xi}_{1,k}^{\phi} \right\|^2 \leq C/B$ holds. Moreover, $\mathbb{E}_k^* \left(\boldsymbol{\Xi}_{1,k}^{\phi} \boldsymbol{\Xi}_{1,k'}^{\phi \mathrm{T}} \right) = 0$ for all $k \neq k'$, so $\mathbb{E}^* \left(\boldsymbol{\Xi}_{1,k}^{\phi} \boldsymbol{\Xi}_{1,k'}^{\phi \mathrm{T}} \right) = 0$ for all $k \neq k'$. We then show that

$$\sup_{k \geq k_n + 1} \mathbb{E}^* \left\| \boldsymbol{\Xi}_{2,k}^{\phi} \right\|^2 = O_p \left(\frac{1}{B h_n^2} \right), \quad \sup_{k \geq k_n + 1} \mathbb{E}^* \left\| \boldsymbol{\Xi}_{3,k}^{\phi} \right\|^2 = O_p \left(\frac{1}{B h_n^2} \right)$$

and

$$\sup_{k,k' \geq k_n + 1, k \neq k'} \left\| \mathbb{E}^* \boldsymbol{\Xi}_{2,k}^{\phi} \boldsymbol{\Xi}_{2,k'}^{\phi \mathrm{T}} \right\| = O_p \left(\frac{\sqrt{\log n}}{B^2 h_n^2} \right), \quad \sup_{k,k' \geq k_n + 1, k \neq k'} \left\| \mathbb{E}^* \boldsymbol{\Xi}_{3,k}^{\phi} \boldsymbol{\Xi}_{3,k'}^{\phi \mathrm{T}} \right\| = O_p \left(\frac{\sqrt{\log n}}{B^2 h_n^2} \right).$$

We will only show the results for $\boldsymbol{\Xi}_{2,k}^{\phi}$. The results for $\boldsymbol{\Xi}_{3,k}^{\phi}$ can be similarly proved. For the first result, according to the proof of Lemma 2, we note that with probability going to 1,

$$\begin{aligned}
\mathbb{E}^*\left\|\boldsymbol{\Xi}_{2,k}^{\phi}\right\|^2 &\leq \frac{1}{B^2}\sum_{j=1}^{B}\sum_{l\neq j}^{B}\mathbb{E}^*\left(\left\|\frac{\mathbf{X}_{i_{k,j}}^{\phi}\mathbf{X}_{i_{k,l}}^{\phi\mathrm{T}}}{A_{n,1,i_{k,j},k}A_{n,1,i_{k,l},k}}\right\|\left|\left(A_{n,1,i_{k,j},k}^{\mathfrak{I}}-A_{n,1,i_{k,j},k}\right)\left(A_{n,1,i_{k,l},k}^{\mathfrak{I}}-A_{n,1,i_{k,l},k}\right)\right|\right)\\
&+\frac{1}{B^2}\sum_{j=1}^{B}\mathbb{E}^*\left(\left\|\frac{\mathbf{X}_{i_{k,j}}^{\phi}\mathbf{X}_{i_{k,j}}^{\phi\mathrm{T}}}{A_{n,1,i_{k,j},k}^2}\right\|\left(A_{n,1,i_{k,j},k}^{\mathfrak{I}}-A_{n,1,i_{k,j},k}\right)^2\right)\\
&\leq \frac{C}{B^2}\sum_{j=1}^{B-1}\sum_{l=j+1}^{B}\frac{1}{Bh_n^2}+\frac{C}{B^2}\sum_{j=1}^{B}\frac{1}{Bh_n^2}\leq\frac{C}{Bh_n^2}.
\end{aligned}$$

The derivation of the second result is more complicated. Without loss of generality, we assume that $\boldsymbol{\Xi}_{2,k}^{\phi}$ is one-dimensional and $k<k'$. Then $\mathbb{E}^*\boldsymbol{\Xi}_{2,k}^{\phi}\boldsymbol{\Xi}_{2,k'}^{\phi}=\mathbb{E}^*\left(\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}\left(\mathbb{E}_{k'}^*\boldsymbol{\Xi}_{2,k'}^{\phi}\right)\right)$. We first look at $\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}$ for general k. We have that

$$\begin{aligned}
&\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}\\
&=\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}_k^*\left[\frac{\mathbf{X}_{i_{k,j}}^{\phi}}{A_{n,1}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right)}\mathbb{E}_k^*\left\{A_{n,y}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right|\mathfrak{I}_{B,k}\right)-A_{n,y}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right)\left|\, j\right\}\right]\\
&=\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}_k^*\left[\frac{\mathbf{X}_{i_{k,j}}^{\phi}}{A_{n,1}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right)}\left\{\mathbb{E}_k^*\left\{\frac{1}{B}\sum_{l=1}^{B}K_{h_n}\left(z_{i_{k,j},k}-z_{i_{k,l},k}\right)y_{i_{k,l}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)y_l\right|\, j\right\}\right\}\right],
\end{aligned}$$

Obviously, for $l\neq j$, we have that $\mathbb{E}_k^*\left\{K_{h_n}\left(z_{i_{k,j},k}-z_{i_{k,l},k}\right)y_{i_{k,l}}\left|\, j\right.\right\}=\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)$. So

$$\begin{aligned}
&\mathbb{E}_k^*\left\{\frac{1}{B}\sum_{l=1}^{B}K_{h_n}\left(z_{i_{k,j},k}-z_{i_{k,l},k}\right)y_{i_{k,l}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)y_l\right|\, j\Bigg\}\\
&=\frac{1}{B}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)y_l\right).
\end{aligned}$$

So

$$\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}=\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}_k^*\left(\frac{1}{B}\frac{\mathbf{X}_{i_{k,j}}^{\phi}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)y_l\right)}{A_{n,1}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right)}\right)$$

Now define $z_i^{\star}=X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}^{\star}$, we have that with probability going to 1, there holds

$$\begin{aligned}
&\left|\frac{\mathbf{X}_{i_{k,j}}^{\phi}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j}}^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_{i_{k,j},k},\boldsymbol{\beta}_k\right)}-\frac{\mathbf{X}_{i_{k,j}}^{\phi}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j}}^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_{i_{k,j}}^{\star},\boldsymbol{\beta}^{\star}\right)}\right|\\
&\leq C\left\|\Delta\boldsymbol{\beta}_k\right\|,
\end{aligned}$$

Then

$$\left|\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}-\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}_k^*\left(\frac{1}{B}\frac{\mathbf{X}_{i_{k,j}}^{\phi}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j}}^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_{i_{k,j}}^{\star},\boldsymbol{\beta}^{\star}\right)}\right)\right|\leq\frac{C\left\|\Delta\boldsymbol{\beta}_k\right\|}{B},$$

which is equivalent to

$$\left|\mathbb{E}_k^*\boldsymbol{\Xi}_{2,k}^{\phi}-\frac{1}{nB}\sum_{i=1}^{n}\left(\frac{\mathbf{X}_i^{\phi}\left(K\left(0\right)y_i-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_i^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)}\right)\right|\leq\frac{C\left\|\Delta\boldsymbol{\beta}_k\right\|}{B},$$

Based on such result, we have that

$$\begin{aligned}&\left|\mathbb{E}_k^*\left(\boldsymbol{\Xi}_{2,k}^{\phi}\left(\mathbb{E}_{k'}^*\boldsymbol{\Xi}_{2,k'}^{\phi}\right)\right)-\mathbb{E}_k^*\left(\boldsymbol{\Xi}_{2,k}^{\phi}\frac{1}{nB}\sum_{i=1}^{n}\left(\frac{\mathbf{X}_i^{\phi}\left(K\left(0\right)y_i-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_i^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)}\right)\right)\right|\\&\leq C\mathbb{E}_k^*\left(\left|\boldsymbol{\Xi}_{2,k}^{\phi}\right|\left\|\Delta\boldsymbol{\beta}_{k'}\right\|\right)/B\leq C\sqrt{\mathbb{E}_k^*\left|\boldsymbol{\Xi}_{2,k}^{\phi}\right|^2}\sqrt{\mathbb{E}_k^*\left\|\Delta\boldsymbol{\beta}_{k'}\right\|^2}/B\leq\frac{C\sqrt{\log n}}{B^2h_n^2}\end{aligned}$$

uniformly for all k when $k\geq k_n+1$. On the other side,

$$\begin{aligned}&\mathbb{E}_k^*\left(\boldsymbol{\Xi}_{2,k}^{\phi}\frac{1}{nB}\sum_{i=1}^{n}\left(\frac{\mathbf{X}_i^{\phi}\left(K\left(0\right)y_i-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_i^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)}\right)\right)\\&=\frac{1}{nB}\sum_{i=1}^{n}\left(\frac{\mathbf{X}_i^{\phi}\left(K\left(0\right)y_i-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_i^{\star}-z_l^{\star}\right)y_l\right)}{A_{n,1}\left(z_i^{\star},\boldsymbol{\beta}^{\star}\right)}\right)\mathbb{E}_k^*\left(\frac{1}{B}\left(K\left(0\right)y_{i_{k,j}}-\frac{1}{n}\sum_{l=1}^{n}K_{h_n}\left(z_{i_{k,j},k}-z_{l,k}\right)y_l\right)\right)\\&=O_p\left(\frac{1}{B^2}\right)\end{aligned}$$

uniformly for all k. This proves the desired result.

Now denote $\widetilde{k}=\left[-\log\left(n\right)/\log\left(1-\delta\underline{\lambda}_{\Lambda}/8\right)\right]$, so $k^*=k_n+\widetilde{k}$. We have that

$$\begin{aligned}&\Delta\boldsymbol{\beta}_{k^*+1+t}\\&=\left(I-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right)^{t+\widetilde{k}}\Delta\boldsymbol{\beta}_{k_n+1}+\delta\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right)^{t+\widetilde{k}-1-k}\Omega_{k_n+1+k}^{\phi}\\&-\delta\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right)^{t+\widetilde{k}-1-k}\boldsymbol{\xi}_n^{\phi}-\delta\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}\right)\right)^{t+\widetilde{k}-1-k}\left(\boldsymbol{\Xi}_{1,k_n+1+k}^{\phi}+\boldsymbol{\Xi}_{2,k_n+1+k}^{\phi}-\boldsymbol{\Xi}_{3,k_n+1+k}^{\phi}\right).\end{aligned}$$

So

$$\frac{1}{T}\sum_{t=1}^{T}\Delta\boldsymbol{\beta}_{k^*+1+t}=\frac{1}{T}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}}\Delta\boldsymbol{\beta}_{k_n+1}+\frac{\delta}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\Omega^\phi_{k_n+1+k}$$
$$-\Lambda_\phi^{-1}\left(\boldsymbol{\beta}^\star\right)\boldsymbol{\xi}^\phi_n-\frac{1}{T}\sum_{t=1}^{T}\left(\delta\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{k}-\Lambda_\phi^{-1}\left(\boldsymbol{\beta}^\star\right)\right)\boldsymbol{\xi}^\phi_n$$
$$-\frac{\delta}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\left(\boldsymbol{\Xi}^\phi_{1,k_n+1+k}+\boldsymbol{\Xi}^\phi_{2,k_n+1+k}-\boldsymbol{\Xi}^\phi_{3,k_n+1+k}\right).$$

We look at the above terms separately. We have that

$$\mathbb{E}^*\left\|\frac{1}{T}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}}\Delta\boldsymbol{\beta}_{k_n+1}\right\|\leq\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{\widetilde{k}}\frac{1}{T}\sum_{t=1}^{T}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{t}\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k_n+1}\right\|$$
$$\leq C\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{\widetilde{k}}\mathbb{E}^*\left\|\Delta\boldsymbol{\beta}_{k_n+1}\right\|=O_p\left(n^{-1}\right),$$

$$\mathbb{E}^*\left\|\frac{\delta}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\Omega^\phi_{k_n+1+k}\right\|\leq\frac{\delta}{T}\sum_{t=1}^{T}\sum_{k=0}^{\infty}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{k}\mathbb{E}^*\left\|\Omega^\phi_{k_n+1+k}\right\|$$
$$\leq C\sup_{k\geq k_n+1}\mathbb{E}^*\left(\left\|\Omega^\phi_k\right\|\right)=o_p\left(n^{-1/2}\right),$$

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\delta\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{k}-\Lambda_\phi^{-1}\left(\boldsymbol{\beta}^\star\right)\right)\boldsymbol{\xi}^\phi_n\right\|=\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\delta\sum_{k=t+\widetilde{k}}^{\infty}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{k}\right)\boldsymbol{\xi}^\phi_n\right\|$$
$$\leq C\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{\widetilde{k}+1}\left\|\boldsymbol{\xi}^\phi_n\right\|=o_p\left(n^{-1/2}\right).$$

We finally look at the last term. We will focus on $\frac{\delta}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\boldsymbol{\Xi}^\phi_{2,k_n+1+k}$ only, because verifying the remaining terms can be done similarly. Without loss of generality, we again assume that $\boldsymbol{\Xi}^\phi_{2,k_n+1+k}$ is one-dimensional. We note that

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\boldsymbol{\Xi}^\phi_{2,k_n+1+k}$$
$$=\frac{1}{T}\sum_{t=1}^{T}\sum_{t'=1}^{t}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t'-1}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}+T-t}+\frac{1}{T}\sum_{l=1}^{\widetilde{k}-1}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+l-1}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}.$$

We have that

$$\begin{aligned}
&\mathbb{E}^*\left(\frac{1}{T}\sum_{l=1}^{\widetilde{k}-1}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+l-1}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}\right)^2\\
&=\mathbb{E}^*\left(\frac{1}{T}\sum_{l=1}^{\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{l}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t-1}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}\right)^2\\
&\leq\frac{1}{T^2}\sum_{t=1}^{T}\sum_{l=1}^{T}\sum_{t'=1}^{t}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{t'-1}\sum_{l'=1}^{l}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{l'-1}\mathbb{E}^*\left(\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}+T-t}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}+T-l}\right)\\
&=O_p\left(\frac{1}{TBh_n^2}+\frac{\sqrt{\log n}}{B^2h_n^2}\right).
\end{aligned}$$

On the other side, we have that

$$\begin{aligned}
&\mathbb{E}^*\left(\frac{1}{T}\sum_{l=1}^{\widetilde{k}-1}\sum_{t=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+l-1}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}\right)^2\\
&=\frac{1}{T^2}\sum_{l=1}^{\widetilde{k}-1}\sum_{l'=1}^{\widetilde{k}-1}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+t'+l+l'-2}\mathbb{E}^*\left(\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}\boldsymbol{\Xi}^\phi_{2,k_n+\widetilde{k}-l}\right)\\
&\leq\frac{C}{T^2}\left(\sum_{l=1}^{\infty}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{l}\right)^4\sup_{k,k'}\left|\mathbb{E}^*\left(\boldsymbol{\Xi}^\phi_{2,k_n+k}\boldsymbol{\Xi}^\phi_{2,k_n+k'}\right)\right|\\
&=O_p\left(\frac{1}{T^2Bh_n^2}\right)
\end{aligned}$$

This implies that

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{k=0}^{t+\widetilde{k}-1}\left(I-\delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{t+\widetilde{k}-1-k}\boldsymbol{\Xi}^\phi_{2,k_n+1+k}=O_{\mathbf{P}}\left(\frac{1}{\sqrt{TBh_n^2}}+\frac{\log^{1/4}(n)}{Bh_n}\right).$$

This proves the result. □

Proof of Theorem 2.3

Proof. To prove the result, it remains to show that

$$P\left(\mathbb{P}^*\lim_{R\to\infty}\widetilde{\Sigma}^\phi_{\boldsymbol{\beta}}=\widehat{\Sigma}^\phi_{\boldsymbol{\beta}}\right)\to 1,$$

where $\widehat{\Sigma}_{\boldsymbol{\beta}}^{\phi}$ is the full-sample-based covariance matrix estimator prposed in Khan et al. (2023). In particular, define

$$\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi}=\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}_i\left(1-\widehat{G}_i\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)\left(\mathbf{X}_i^{\phi}-\widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)\right)^{\mathrm{T}}\right),$$

and

$$\widehat{\Lambda}_{\phi}\left(\widehat{\boldsymbol{\beta}}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\overline{\boldsymbol{\beta}}\right)\middle|\overline{\boldsymbol{\beta}}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}},$$

where

$$\widehat{G}_i=\frac{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)y_j}{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)},\ \widehat{\mathbb{E}}\left(\mathbf{X}_i^{\phi}\middle|\widehat{z}_i\right)=\frac{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)\mathbf{X}_j^{\phi}}{\sum_{j=1}^{n}K_{h_n}\left(\widehat{z}_i-\widehat{z}_j\right)},$$

and $\widehat{z}_i=X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\overline{\boldsymbol{\beta}}$. Then $\widehat{\Sigma}_{\boldsymbol{\beta}}^{\phi}$ is defined by $\widehat{\Sigma}_{\boldsymbol{\beta}}^{\phi}=\widehat{\Lambda}_{\phi}^{-1}\left(\overline{\boldsymbol{\beta}}\right)\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi}\left(\widehat{\Lambda}_{\phi}^{-1}\left(\overline{\boldsymbol{\beta}}\right)\right)^{\mathrm{T}}$. So we only need to show that, with probability going to 1,

$$\frac{1}{R}\sum_{r=1}^{R}\widehat{\Lambda}_{\phi}^{r}\left(\overline{\boldsymbol{\beta}}\right)\rightarrow_{\mathbb{P}^*}\widehat{\Lambda}_{\phi}\left(\overline{\boldsymbol{\beta}}\right)$$

and

$$\frac{1}{R}\sum_{r=1}^{R}\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi,r}\rightarrow_{\mathbb{P}^*}\widehat{\Sigma}_{\boldsymbol{\xi}}^{\phi}$$

as R increases to infinity. This can be easily done using the previous proof method.

□

Chapter 3

Quantile Control via Random Forest

3.1 Introduction

Estimating treatment effects in panel data with only one treated unit has attracted a large amount of research attention in applied work. Due to the limitation of the real data, it is not uncommon that the econometrician may fail to observe the key factors that drive the evolution of the outcomes of the treated and untreated units. This motivates the use of the outcomes of the control units in the panel as proxies for these unobserved factors to predict the outcome of the treated unit. Popular methods include synthetic control method (SCM, Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015) and regression control method (RCM, Hsiao et al., 2012; Hsiao and Zhou, 2019)[1]. Collectively, these methods are sometimes known as "synthetic control methods" (Cattaneo et al., 2021) or "counterfactual and synthetic control methods" Chernozhukov et al. (2021b).

Despite their great popularity among empirical researchers, statistical inference for these methods is still an active research area. The first contribution of this paper is on studying robust inference of treatment effects under the SCM framework that accommodates flexible relationship across different units as well as high dimensionality. Compared with the existing methods such as those in Chernozhukov et al. (2021b) and Cattaneo et al. (2021), our proposed method is a more robust approach. In particular, we do not make assumptions over the functional relationship between the outcomes

[1]Since Hsiao et al. (2012) use regression to construct the counteractual control unit, we coin the term "regression control method" in the same spirit as "synthetic control method". Gardeazabal and Vega-Bayo (2017) and Wan et al. (2018) compare the empirical performance of synthetic control method and regression control method using simulations and real datasets, but reach different conclusions.

of the treated and control units, nor do we restrict the behavior of the projection error (such as homoskedasticity). Instead, we only require that there is a stable distributional relationship between the outcomes of the treated unit and the control units, so that information on the quantiles of the treated unit can be inferred from the observations of the control units. Our inference procedure does not depend on the assumption of random assignment of interventions or the symmetry assumption, nor does it require a large number of post-treatment periods. To accommodate such a general model structure, we propose to use a machine learning technique called "Quantile Random Forest" (QRF), also known as "Quantile Regression Forest" (Meinshausen, 2006), to efficiently and robustly estimate the conditional distribution of treatment effects. The proposed method is robust to heteroskedasticity, autocorrelation and various types of model misspecifications. Since our proposed method uses quantile regression and QRF in particular to construct a synthetic counterfactual control unit as well as its relevant quantiles, we call it "Quantile Control Method" (QCM). Our Monte Carlo simulations show that, comparing to methods in the existing literature, prediction intervals via QCM have excellent coverage probability for the treatment effects even in small samples.

Meinshausen (2006) originally proposes the QRF and develops a framework to establish its consistency. Recently, Athey et al. (2019) propose "Generalized Random Forest", which offers an alternative algorithm to estimate the conditional quantile function via Random Forest based on the gradient of the check function as in Koenker and Bassett (1978) and established its asymptotic property for "honest" trees. In our unreported simulations, the performance of generalized random forest is similar to the original QRF by Meinshausen (2006). Therefore, we stick with the latter for the simplicity of its algorithm.[2]

As another contribution of this paper, we formally establish the asymptotic validity of the QRF under the setup of weak dependence and high dimensionality, which nests SCM as a special case but also applies to general high-dimensional time series scenarios. Despite its track record as one of the best predictive algorithms and huge popularity among data scientists and practitioners, the asymptotic theory for Random Forest is still a growing area. Breiman (2004) offers heuristics for the consistency of a simplified version of Random Forest. Biau (2012) formalizes Breiman (2004)'s approach, and provides a proof of consistency based on similar assumptions. Scornet et al. (2015) provide a more general proof of consistency at the cost of imposing an additive regression model. Mentch and

[2]A related literature to the high dimensional quantile regression is the penalized quantile regression approach (He et al., 2013; Wang et al., 2012). For example, Belloni and Chernozhukov (2011) derived an error bound for sparse high dimensional linear quantile regression with the L_1-penalty. Wang et al. (2012) studied such regressions based on nonconvex penalties MCP and SCAD. Various extensions are developed along this direction, see Belloni et al. (2017) for additional discussions on high-dimensional quantile regression methods and related literature.

Hooker (2016) derive the asymptotic distribution of Random Forest by replacing bootstrapping with subsampling and making use of the theory of U-statistic. Wager and Athey (2018) also develop asymptotic normality of random forests, based on a different set of assumptions that require the trees to be "honest" and "regular". Meinshausen (2006) provides a framework to establish the consistency of the QRF. Despite its novelty, there are some theoretical limitations in his framework that limit the potential use of the QRF. For example, in Meinshausen (2006)'s work, the data is required to be iid and fixed-dimensional. This leads to a question of whether the QRF can be applied to the data with weak dependence and high dimensionality. Moreover, when constructing trees, it is assumed that each covariate will be selected as splitpoint with probability bounded from below, which leaves the tree growing procedure a "black box". It is important to answer whether such high-level assumption will actually hold when the proposed algorithm is empirically applied.

In this paper, we address these issues and formally establish the validity of the QRF algorithm for data with weak dependence and high dimensionality. We show that, under the algorithm proposed by Breiman (2001), almost all trees in the forest will choose signal variables with increasing number of times under some sparsity conditions. Our proof first applies the "approximating rectangles" method proposed by Wager and Walther (2016) to show concentration of forest prediction. However, different from Wager and Walther (2016), we do not impose a Guess-and-Check tree structure when analyzing the bias of forest prediction as well as showing the consistency of QRF. Indeed, the Guess-and-Check procedure is designed to screen out noise covariates that are independent of the response variable. While under the time series/panel data setup such as SCM, although some variables do not directly affect the response, they are not necessarily independent of the response. For example, when analyzing Hong Kong's economic growth, Hsiao et al. (2012) find that only Austria, Italy, Korea, Mexico, Norway and Singapore have nonzero impacts, but this does not imply that Hong Kong's economic growth is completely independent of the growth of the US. In fact, as long as the US economy has impacts on the economic growth of the above six countries, such impacts could be transmitted to Hong Kong's economic growth, which leads to correlation between Hong Kong's and the US's economic growth. Above analysis implies that Guess-and-Check procedure may choose too many noise covariates due to their dependence with the response and lead to too many splits along them. To deal with this issue, we follow the original splitting procedure proposed by Breiman (2001). Under some regularity conditions, we show that as long as all the noise variables are conditionally independent of the response variables, and almost all of them contain less "information" compared with the signal variables, then for almost all trees, each signal variable will be selected as the

splitting point with increasing number of times as the tree grows, and hence the consistency of QRF holds. Our results not only facilitate the inference of the treatment effect based on the panel/time series data environment in SCM, but also are applicable to general time series analysis with high dimensionality.

We conduct extensive Monte Carlo experiments to investigate the performance of our method. We compare the proposed method with other methods in the literature. The simulation results show that confidence intervals via QCM have excellent coverage probability for the treatment effects even in small samples, and is robust to the presence of heteroskedasticity, autocorrelation and nonlinear functional forms. Under a variety of DGPs including linear or nonlinear factor models, as well as models free of factor structures with or without sparsity, the proposed QCM prediction intervals enjoy outstanding performance across the board.

The proposed QCM approach provides a useful inferential tool for applied work of policy evaluations. As an illustration, we apply QCM to revisit the example on studying the effect of the economic integration between Hong Kong and mainland China on Hong Kong's economy (Hsiao et al., 2012). QCM can be easily implemented by using forthcoming packages `qcm` in both R and Stata.

3.1.1 Literature Review

A popular way of inference for SCM or RCM relies on design-based placebo test (Abadie et al., 2010; Gardeazabal and Vega-Bayo, 2017). This in-space placebo test is akin to permutation tests used by classical randomization inference when the intervention is randomly assigned, which, however, is not a probable setting especially in the contexts with aggregate units. Hahn and Shi (2017) point out that the validity of permutation tests depends on the symmetry assumption, which may not hold in the case of SCM. Also see Carvalho et al. (2018); Galiani and Quistorff (2017); Firpo and Possebom (2018); Ferman and Pinto (2017) for other work on placebo tests. Another approach of statistical inference focuses on average treatment effect (ATE) for the single treated unit over the entire post-treatment periods, such as Carvalho et al. (2018); Chernozhukov et al. (2018); Li (2020); Shi and Huang (2021). The asymptotic theories require the number of post-treatment periods to be large, which may not be satisfied in empirical work.

For pointwise inference, Fujiki and Hsiao (2015) provide a simple textbook formula for the standard errors and confidence bands of the treatment effects based on a strong assumption of i.i.d. errors.

Xu (2017) uses the interactive fixed effects model to impute treated counterfactuals, and proposes a parametric block bootstrap of the residuals to obtain uncertainty estimates of the average treatment effect on the treated (ATT) based on assumptions of correctly specified parametric model and homoskedastic error terms. Arkhangelsky et al. (2021) combine insights from difference in differences and SCM in a "synthetic difference in differences" estimator and propose a variance estimator based on placebo tests similar to Abadie et al. (2010). This "placebo variance estimator" also relies fundamentally on homoskedasticity across units. In addition, Bayesian approaches have also been adopted to tackle the issue of statistical inference for SCM, such as Amjad et al. (2018); Kim et al. (2020); Pang et al. (2022).

In a recent paper, Chernozhukov et al. (2021b) study the inference of SCM based on the assumption that the underlying model is able to generate a mean-unbiased proxy P_t^0 for the counterfactual outcome of the treated unit in the absence of the policy. They investigate several models regarding the specification of P_t^0, Based on these models, they consider testing hypotheses about the treatment effect θ_t. The proposed method chooses a fine grid of values of θ_t, say, $\{\theta_1^*, \cdots, \theta_G^*\}$. For each candidate value θ_g^*, the mean-unbiased proxy P_t^0 can be estimated based on the model and the corresponding null restricted data. Then a conformal inference of hypothesis testing by permuting blocks of estimated residuals is proposed. In another recent paper, Cattaneo et al. (2021) consider a linear model between the features of the treated unit, a_t, and features of the untreated units and control variables p_t. They consider the linear least square problem $\widehat{\beta} = \arg\min_\beta \sum_{t=1}^{T_0} (a_t - p_t'\beta)^2$, and the predictive interval of the treatment effect are constructed by approximating the uncertainty in $p_T^T(\beta - \widehat{\beta})$ and e_T, where e_T is the projection error. Cattaneo et al. (2021) propose to approximate the uncertainty in $p_T^T(\beta - \widehat{\beta})$ using simulation approximation based on random draws from Gaussian variates with an appropriate estimator of the covariance matrix. Based on different assumptions on the model structure, Cattaneo et al. (2021) discuss model specifications corresponding to iid, stationary, and unit root nonstationary data. Approximating the out-of-sample uncertainty in e_T requires additional strong distributional assumptions. Cattaneo et al. (2021) discuss three different strategies to assess the uncertainty based on progressively stronger restrictions. Such assumptions, "however, are difficult to avoid" (Cattaneo et al., 2021).

Finally, we would like to make a comment on nonstationarity. The current paper focuses on the case where the data is stationary over t. Cattaneo et al. (2021) and Chernozhukov et al. (2021b) discussed models with certain types nonstationarity such as cointegration. For example, Cattaneo et al. (2021) considered the nonstationary case where the pre-treatment outcomes are integrated

process. In such cases, the pre-treatment outcomes are differenced so that stationarity can be achieved. The variance of the (stationary) differenced data can be estimated and random samples can be drawn from a normal distribution with appropriate estimated variance matrix. Note that under appropriate assumptions on the form of nonstationarity, we may transform the nonstationary data into stationary one, and then our proposed method may be applied to the transformed data. However, when the nature of nonstationarity is unknown, such transformations are infeasible and existing methods will generally be invalid[3].

3.1.2 Organization

The rest of this paper is arranged as follows. Section 3.3.2 formally introduces the setup of our problem. Section 3.3 presents the inference procedure for the treatment effects via QCM. Section 3.4 proves the asymptotic consistency of the QCM. Section 3.5 reports Monte Carlo simulations to demonstrate the small-sample properties of QCM. As an empirical illustration of our proposed method, Section 3.6 applies the proposed method to study the effect of the economic integration between Hong Kong and mainland China on Hong Kong's economy. Finally, Section 3.7 concludes. Additional results for simulation and empirical applications, and auxiliary results for theoretical establishment are rearranged to the Supplementary Material to this paper.

3.2 The Model

Suppose that we observe panel data with outcome variables Y_{it} for individuals $i = 1, ..., n+1$ (for example, "regions" in regional policy evaluation, such as countries, states or cities), over periods $t = 1, \cdots, T_0, T_0+1, \cdots, T_0+T_1 :\equiv T$. The time dimension T is divided into two parts: $T_0 + T_1$, where T_0 is the number of pre-treatment periods (from period 1 through period T_0), and T_1 is the number of post-treatment periods (from period $T_0 + 1$ through period $T_0 + T_1$). Without loss of generality, assume that the first individual is the only treated unit, while all other individuals are control units, which form a donor pool[4]. In other words, a policy intervention or treatment happens to the first unit from period T_0+1 through period T_0+T_1, while all other units receive no treatment

[3]Indeed, Chernozhukov et al. (2021b) point out that "both unrestricted patterns of nonstationarity and misspecification is not possible in general. To obtain valid inferences with nonstationary data, one has to either rely on correct specification and consistency or impose assumptions on the particular structure of the non-stationarity, which allow for preprocessing the data to make them stationary."

[4]The case of multiple treated units can be accomodated by applying the same procedure to each treated unit separately.

throughout.

Following Rubin's causal model, denote Y_{it}^1 and Y_{it}^0 as the potential outcomes with and without treatment for individual i in period t. The observed outcome is given by $Y_{it} = d_{it}Y_{it}^1 + (1 - d_{it})Y_{it}^0$, where d_{it} is a dummy indicating the treatment status for unit i in period t. The treatment effect for unit i in period t is defined as $\Delta_{it} = Y_{it}^1 - Y_{it}^0$. Our basic interest is to make period-wise inference on the treatment effect of the first unit based on the data we observe. A fundamental problem of the inference is that we do not simultaneously observe both Y_{1t}^0 and Y_{1t}^1 at the same time. Note that $Y_{1t} = Y_{1t}^0$ for $t \leq T_0$, and $Y_{1t} = Y_{1t}^1$ for $t \geq T_0 + 1$. To make inference on the treatment effects on the first treated unit, we need information on the unobservable Y_{1t}^0 in the post-treatment periods.

Let $Z_{1t}, ..., Z_{pt}$ be covariates that can be used to predict Y_{1t}^0, and $N = n + p$. To ease our notation, we denote $Y_t = Y_{1t}^0$ and $\mathbf{X}_t = (Y_{2t}, \cdots, Y_{n+1,t}, Z_{1t}, \cdots, Z_{pt})^T$. A key operating assumption in the existing literature for treatment effect estimation and inference under the SCM framework is that there exists a cross-sectional relationship between Y_t and $\mathbf{X}_t$. In this paper, we assume that the relationship between Y_t and $\mathbf{X}_t$ is characterized by

$$Y_t | \, \mathbf{X}_t \sim f \left(Y | \, \mathbf{X} \right), t = 1, \cdots, T, \tag{3.1}$$

where $f \left(\cdot | \, \cdot \right)$ is an unknown conditional density function. Equation (3.1) implies that if the treatment never occurs, the distribution of the outcomes of the treated unit conditional on all the control units will remain stable throughout time.

Now we make some comparisons between our setup in (3.1) and the setup in the conventional SCM. The conventional SCM assumes that there is a linear relationship between the outcomes of the treated and control units (Abadie et al., 2010; Hsiao et al., 2012; Amjad et al., 2018; Cattaneo et al., 2021), that is,

$$Y_t = \mathbf{X}_t^T W_0 + \varepsilon_t, \tag{3.2}$$

where W_0 is the (pseudo) true linear projection parameter and ε_t is the error term. To predict the (conditional) mean of Y_t, it remains to estimate W_0; see Cattaneo et al. (2021) for an excellent review for the estimation methods of W_0. While such linear setup makes the SCM easy to implement and interpret, it misses some important information. On the one side, some latent factors may have asymmetric impacts on different individuals, rendering (3.2) misspecified. Even though in many situations we may view linear synthetic control as a reasonable approximation for the true

process[5], distributional information on the quantiles is inevitably ignored when we focus on the mean prediction. For example, a region is more likely to experience negative shocks when its neighboring regions are in economic downturns, indicating that the distribution of the region's economic growth is shifted leftwards and may have a thick left tail when we observe negative growth rates of its neighboring regions. If we focus on conditional mean only, such useful information might be missed.

Comparatively, our setup (3.1) is a natural generalization of the conventional SCM. It demonstrates that the distributional information of the outcome of the first unit without treatment can be deduced from the observations of the outcomes of the control units. Apart from stable conditional distribution, it does not impose any model structure on the cross-sectional dependence between the treated and control units. What mainly distinguishes our setup (3.1) from the setup of the conventional SCM is that (3.1) goes beyond conditional mean and provides information over the quantiles, so the distribution of Y_t is completely determined after we have observed the outcomes of the units in the donor pool. Given such distributional information, we can construct the prediction interval for Y_t, which we will show later can be used to make inference on the period-wise treatment effects. We also point out that the specification in equation (3.1) is in fact very general such that it encompasses both factor and non-factor based models. In particular, factor models as used by Abadie et al. (2010) and Hsiao et al. (2012) are special cases of (3.1).

Remark 3.1. *We make two additional comments on the setup (3.1). First of all, although (3.1) assumes that the conditional distribution is stable throughout time, we do not rule out time dependence. See Assumption 3.1 and Assumption 3.2 in section 3.4 below. Second, the conventional SCM estimator for the treatment effect can be expressed as a functional of estimated* $f(Y|\mathbf{X})$*. In particular, if we have an estimator for* $f(Y|\mathbf{X})$*, denoted as* $\widehat{f}(Y|\mathbf{X})$*, then the point estimate of treatment effect is given by* $\widehat{\Delta}_{1t} = Y_{1t} - \widehat{Y}^0_{1t}$*, where* $\widehat{Y}^0_{1t} = \int y\widehat{f}(y|\mathbf{X})\,dy$*. In addition, based on the point estimate of the treatment effect* $\widehat{\Delta}_{1t}$*, the average treatment effect on the first unit from period* T_0+1 *to period* T_0+T_1 *can be estimated as* $\widehat{\Delta}_1 \equiv \frac{1}{T_1}\sum_{t=T_0+1}^{T_0+T_1}\widehat{\Delta}_{1t}$*.*

3.3 The Quantile Control Method

In the case of linear models specified in (3.2), the original papers of both SCM (Abadie and Gardeazabal, 2003) and RCM (Hsiao et al., 2012) rely on informal inference. In many empirical applications,

[5]For example, Cattaneo et al. (2021) define W_0 and ε_t as the pseudo true linear projection parameter and residual that satisfy $W_0 = \arg\min_W \mathbb{E}((Y_t - \mathbf{X}_t^T W)^2|\mathcal{H})$, where $\mathcal{H}$ is an information set, and $\varepsilon_t = Y_t - \mathbf{X}_t^T W_0$.

the predicted outcome for the first unit before treatment closely tracks its observed outcome, which lends support to the trustworthiness of the imputed counterfactual outcome for the first unit after treatment. Typically, a "gap graph" is drawn to reveal the divergence between the observed and counterfactual outcomes after treatment (in contrast to their closeness before treatment), as a way to showcase the presumably significant treatment effects. However, this type of informal inference is unsatisfactory. First, the pre-treatment in-sample fit may not be a reliable indicator of the model's ability to predict future data that it has not yet seen, which is widely known as "overfitting" in the machine learning literature. Second, it is possible that despite an imperfect pre-treatment fit, the estimated treatment effects are still significant. Consequently, requiring perfect pre-treatment fit unnecessarily restricts the applicability of synthetic control methods in applied work[6]. Both issues can be remedied if we could provide valid prediction intervals for the treatment effects. For example, even if the predicted outcome for the first unit does not track the actual outcome before the treatment very well, the treatment effects might nevertheless be significant if the prediction intervals do not contain zero for some periods after the treatment, as these confidence intervals have already taken into account the uncertainty from the imperfect pre-treatment fit.

Motivated by our model setup (3.1), in this paper we propose to estimate the conditional quantiles/distribution of unobserved Y_{1t}^0 given the observation of $Y_{2t}^0, \cdots, Y_{n+1,t}^0, Z_{1t}, \cdots, Z_{pt}$, and then construct the prediction intervals for the treatment effects via quantile regression (Koenker and Bassett, 1978; Koenker, 2005). The usefulness of quantile regression as a way to construct prediction intervals have long been recognized and proven in the statistics literature (e.g. Zhou and Portnoy, 1996; Koenker, 2005).

3.3.1 Prediction Intervals Based on Quantile Regression

We first introduce the general framework based on which we construct the prediction intervals for the treatment effect. Recall that we denote Y_{1t}^0 as Y_t and $(Y_{2t}, \cdots, Y_{n+1,t}, Z_{1t}, \cdots, Z_{pt})^T$ as $\mathbf{X}_t$. To construct a point-wise prediction interval of Δ_{1t} for $t \geq T_0 + 1$ with a confidence level $(1 - 2\alpha)$, we start with the α and $(1 - \alpha)$ quantiles of the counterfactual outcome Y_t. Denote $Q_{Y_t}(\alpha | \mathbf{X}_t)$ and

[6]Ben-Michael et al. (2021) propose to de-bias the synthetic control estimator with imperfect pre-treatment fit via ridge regression. However, Ferman and Pinto (2021) warn that when the pre-treatment fit is imperfect, synthetic control estimators are generally biased if treatment assignment is correlated with time-varying unobserved confounders, even when the number of pre-treatment periods goes to infinity. Nevertheless, Ferman (2021) shows that this bias goes away if both the number of pre-treatment periods and the number of control units tend to infinity.

$Q_{Y_t}(1-\alpha|\mathbf{X}_t)$ as the α and $(1-\alpha)$ conditional quantiles of Y_t, respectively. Then we have

$$P\left(Q_{Y_t}(\alpha|\mathbf{X}_t) \leq Y_t \leq Q_{Y_t}(1-\alpha|\mathbf{X}_t)|\mathbf{X}_t\right) = 1-2\alpha.$$

Since $Y_t = Y_{1t}^0$, we have that $\Delta_{1t} = Y_{1t}^1 - Y_t$, and

$$P\left(Q_{Y_t}(\alpha|\mathbf{X}_t) \leq Y_{1t}^1 - \Delta_{1t} \leq Q_{Y_t}(1-\alpha|\mathbf{X}_t)\Big|\mathbf{X}_t\right) = 1-2\alpha,$$

which is equivalent to:

$$P\left(Y_{1t} - Q_{Y_t}(1-\alpha|\mathbf{X}_t) \leq \Delta_{1t} \leq Y_{1t} - Q_{Y_t}(\alpha|\mathbf{X}_t)|\mathbf{X}_t\right) = 1-2\alpha. \tag{3.3}$$

(3.3) provides a theoretical prediction interval with confidence level $(1-2\alpha)$ for the treatment effect Δ_{1t}. If $Q_{Y_t}(\alpha|\mathbf{X}_t)$ and $Q_{Y_t}(1-\alpha|\mathbf{X}_t)$ were known, the prediction interval for Δ_{1t} can be readily constructed. In practice, suppose that we obtain consistent estimators $\widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t)$ and $\widehat{Q}_{Y_t}(1-\alpha|\mathbf{X}_t)$, then

$$P\left(\widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t) \leq Y_t \leq \widehat{Q}_{Y_t}(1-\alpha|\mathbf{X}_t)\Big|\mathbf{X}_t\right) \to_p 1-2\alpha$$

and an asymptotic $(1-2\alpha)$ prediction interval for Δ_{1t} can be constructed by

$$\left[Y_{1t} - \widehat{Q}_{Y_t}(1-\alpha|\mathbf{X}_t), Y_{1t} - \widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t)\right]. \tag{3.4}$$

In the special case where $f(\cdot|\cdot)$ in (3.1) belongs to some parametric model family, we may use a parametric (usually linear) quantile regression to estimate the α- and $(1-\alpha)$-th conditional quantiles of Y_t. But a parametric quantile relationship may be generally misspecified in practice. Indeed, even if $\mathbb{E}(Y_t|\mathbf{X}_t)$ is a linear function of $\mathbf{X}_t$ as in (3.2), when the error term has heteroskedasticity and/or autocorrelation of unknown forms, the linear quantile regression could be inappropriate and the estimated quantiles will be poor. Furthermore, Monte Carlo simulations (unreported to save space) show that even under correct model specification, prediction intervals based on linear quantile regressions converge too slowly to have satisfactory coverage probability in finite samples. Also see Chernozhukov et al. (2021a) for related discussion on the "plug-in" approaches.

To overcome these theoretical and practical issues based on traditional quantile regressions, we

consider the most general setup as specified in (3.1) and propose a method based on Random Forest. Suppose that we can obtain a (uniformly) consistent estimator for the conditional distribution function $F(Y|\mathbf{X})$, denoted as $\widehat{F}(Y|\mathbf{X})$, then given any observation $\mathbf{X}_t$ in the post-treatment period, the conditional α-quantile of Y_t given $\mathbf{X}_t$ can be constructed as

$$\widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t) = \inf\left\{y : \widehat{F}(y|\,\mathbf{X}_t) \geq \alpha\right\}. \tag{3.5}$$

Then we can plug (3.5) into (3.4) and the prediction interval is constructed. The above analysis implies that all the problems now boil down to how we can construct a consistent estimator for $F(Y|\mathbf{X})$. When N is very small and T_0 is large, nonparametric estimation procedure such as kernel method can be readily applied. However, when using SCM for policy evaluations, the number of pre-treatment periods is often moderate, and at the same time, the number of control units is relatively large. Both of the problems make conventional nonparametric procedure infeasible in the applications. To address the above concerns, we propose to use the Quantile Random Forest (QRF, Meinshausen, 2006) for the construction of $\widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t)$, which will be discussed in detail in the next subsection.

3.3.2 Quantile Random Forest

Before we discuss the QRF method, we first briefly introduce the regression trees and Random Forest. Given a training set $\{Y_t, \mathbf{X}_t\}_{t=1}^{T_0}$, the regression tree and Random Forest both aim to predict $\mathbb{E}(Y|\,\mathbf{X})$ at the test point $\mathbf{X}$. Denote the N-dimensional feature space of $\mathbf{X}_t$ as $\chi \subseteq \mathbb{R}^N$, the popular CART (Classification and Regression Tree) algorithm grows a tree by recursive binary axis-parallel partition of χ. In the case of regression trees for continuous response Y_t, the partitioning rule at each node is to minimize the mean squared errors by selecting the best splitting variable and splitting position (also known as "cut"). In particular, given any node (rectangle) $R = \bigotimes_{i=1}^{N}[r_i^-, r_i^+] \subseteq \chi$, define $r_{i,\lambda} = \lambda r_i^- + (1-\lambda)\, r_i^+$ and

$$\begin{aligned}\widehat{\mathcal{I}(R,i,\lambda)} = \widehat{\sigma}^2(Y|\,R) &- \frac{\#(R \cap \{X_i < r_{i,\lambda}\})}{\#R}\widehat{\sigma}^2(Y|\,R \cap \{X_i < r_{i,\lambda}\}) \\ &- \frac{\#(R \cap \{X_i \geq r_{i,\lambda}\})}{\#R}\widehat{\sigma}^2(Y|\,R \cap \{X_i \geq r_{i,\lambda}\}),\end{aligned} \tag{3.6}$$

where $\widehat{\sigma}^2(Y|\,R) = \widehat{Y}^2(R) - (\widehat{Y}(R))^2$, $\widehat{Y}(R) = \frac{1}{\#R}\sum_{\mathbf{X}_t \in R} Y_t$, $\widehat{Y}^2(R) = \frac{1}{\#R}\sum_{\mathbf{X}_t \in R} Y_t^2$, and $\#R = |\{t : \mathbf{X}_t \in R\}|$. Let $(\widehat{i}, \widehat{\lambda}) = \arg\max_{1 \leq i \leq N, \lambda} \widehat{\mathcal{I}(R,i,\lambda)}$, then R is split into two subsets $R_L =$

$R \cap \{X_{\widehat{i}} < r_{\widehat{i},\widehat{\lambda}}\}$ and $R_R = R \cap \{X_{\widehat{i}} \geq r_{\widehat{i},\widehat{\lambda}}\}$. After many rounds of cuts, χ is partitioned into a number of hyper-rectangles known as "terminal nodes" or "leaves". The prediction by a single tree at $\mathbf{X}$ is just the average of all Y_t's falling into the same leaf as $\mathbf{X}$.

In general, the prediction based on a single tree is a discontinuous function of the data and the new test point, which has large variance. To improve the prediction accuracy, Breiman (2001) proposes Random Forest as an ensemble learning algorithm by injecting randomness into the tree-building process, and then taking the average of the resulting randomized trees. Specifically, M resamples based on bootstrap are obtained from the original training data to grow M trees. To further diminish the correlations among these trees, random feature selection is conducted. In particular, at each node of an individual tree, only m_{try} features are randomly chosen out of all N features as candidate splitting variables, where m_{try} is a tuning parameter[7]. Since the panel data evolves over the time dimension, we skip the bootstrap procedure and simply use the original sample to preserve its time series nature[8].

Due to the randomization in growing the forest, each tree in the forest can be denoted as $T(\theta_m)$, $m = 1, ..., M$, where θ_m is an iid random object characterizing the randomness in the m-th tree[9]. For each tree, its leaves are determined by the training data as well as the randomness during partition described by θ_m. For any test point $\mathbf{X} \in \chi$, there is a unique leaf that contains $\mathbf{X}$. We denote such leaf as $R(\mathbf{X}, \theta_m)$. Basically, $R(\mathbf{X}, \theta_m)$ contains all the feature vectors that are close to $\mathbf{X}$ under the m-th tree. Then the prediction of $\mathbb{E}(Y | \mathbf{X})$ based on tree $T(\theta_m)$ is obtained by averaging over the observed values of Y_t in the leaf $R(\mathbf{X}, \theta_m)$. Alternatively, we can view the prediction as a weighted average of all Y_t in the sample, where observations in the same leaf as $\mathbf{X}$ receive equal weights that sum to one, and all other observations receive zero weights. Thus, the prediction based on tree $T(\theta_m)$ can be written as

$$\widehat{m}(\mathbf{X}, \theta_m) = \sum_{t=1}^{T_0} w_t(\mathbf{X}, \theta_m) Y_t, \tag{3.7}$$

where the weight $w_t(\mathbf{X}, \theta_m)$ is given by

$$w_t(\mathbf{X}, \theta_m) = \frac{\mathbf{1}(\mathbf{X}_t \in R(\mathbf{X}, \theta_m)}{\#\{s : \mathbf{X}_s \in R(\mathbf{X}, \theta_m)\}}, \tag{3.8}$$

[7]For regression forest, m_{try} is by default set to be $[N/3]$ in R or Matlab.

[8]In our unreported simulations, we find that the performance of using the original sample is similar or even slightly better than using bootstrap samples. This is consistent with Breiman (2004)'s claim that "omitting the bootstrapping of the training set has very little effect on the error rate".

[9]For example, when the bootstrap procedure is omitted, the randomness in each tree lies only in the procedure of feature selection, then θ_m is a collection of random sets of indices, indicating the sets of covariates that are considered for maximization of (3.6) at each parent node in the tree.

where $\mathbf{1}(\cdot)$ is the indicator function. Finally, Random Forest predicts $\mathbb{E}(Y|\mathbf{X})$ based on the combination of the prediction results from the M trees. Define $w_t(\mathbf{X})$ as the average of $w_t(\mathbf{X}, \theta_m)$ over these trees,

$$w_t(\mathbf{X}) = \frac{1}{M}\sum_{m=1}^{M} w_t(\mathbf{X}, \theta_m). \tag{3.9}$$

The prediction from the Random Forest is then given by

$$\widehat{m}(\mathbf{X}) = \sum_{t=1}^{T_0} w_t(\mathbf{X}) Y_t. \tag{3.10}$$

Remark 3.2. *Based on Random Forest, given the pre-treatment training set* $\{Y_s, \mathbf{X}_s\}_{s=1}^{T_0}$ *and post-treatment observations of* $\mathbf{X}_t$, $t \geq T_0 + 1$, *the prediction of the conditional mean of the first unit without treatment is given by* $\widehat{Y}_t = \widehat{m}(\mathbf{X}_t)$. *Then the point estimate of the treatment effect of period* t *is given by* $\widehat{\Delta}_{1t} = Y_{1t} - \widehat{Y}_t = Y_{1t} - \widehat{m}(\mathbf{X}_t), t \geq T_0 + 1$.

Next we consider QRF, which is also known as "Quantile Regression Forest" (Meinshausen, 2006), as an efficient and robust implementation of quantile regression. Inheriting from all the merits of Random Forest algorithm, QRF is free of restrictive assumptions on functional form and thus is robust to heteroskedasticity and/or autocorrelation of unknown forms as well as various types of model misspecifications. Basically, QRF aims to estimate $Q_Y(\alpha|\mathbf{X})$ via Random Forest. Assuming that the response variable Y_t is continuous, as we showed in (3.5), the estimator of $Q_Y(\alpha|\mathbf{X})$ can be constructed as the inverse of the estimated conditional distribution function $F(y|\mathbf{X})$. Note that for any $y \in \mathbb{R}$, $F(y|\mathbf{X})$ can be written as

$$F(y|\mathbf{X}) = P(Y \leq y|\mathbf{X}) = \mathbb{E}\mathbf{1}(Y \leq y|\mathbf{X}),$$

where $\mathbf{1}(\cdot|\mathbf{X})$ is the indicator function conditional on the observation $\mathbf{X}$. Similar to how we predict $\mathbb{E}(Y|\mathbf{X})$ by a weighted combination of Y_t's in the Random Forest algorithm, to predict the conditional expectation $\mathbb{E}\mathbf{1}(Y \leq y|\mathbf{X})$, we can similarly define the estimator by the weighted average over the observations of $\mathbf{1}(Y \leq y)$ given $\mathbf{X}$:

$$\widehat{F}(y|\mathbf{X}) = \sum_{t=1}^{T_0} w_t(\mathbf{X})\mathbf{1}(Y_t \leq y), \tag{3.11}$$

where the weights $w_t(\mathbf{X})$'s are the same as those for Random Forest defined in (3.9). The QRF estimator $\widehat{Q}_Y(\alpha|\mathbf{X})$ is obtained by plugging $\widehat{F}(y|\mathbf{X})$ to (3.5), based on which the estimated prediction

interval can be constructed based on (3.4).

We summarize the proposed QCM Algorithm for convenience of application.

Algorithm QCM

Data: $\{Y_{it}\}$, where $i = 1$ is the treated unit, and $i = 2, ..., n+1$ are control units; $Z_{jt}, j = 1, \cdots, p$, are covariates; $t = 1, ..., T_0$, are pretreatment periods, and $t = T_0+1, ..., T_0+T_1$, are posttreatment periods. Denote $Y_t \equiv Y_{1t}$ and $\mathbf{X}_t \equiv (Y_{2t}, ..., Y_{n+1,t}, Z_{1t}, \cdots, Z_{pt})'$.

Algorithm: $(1-\alpha)$ Prediction intervals of treatment effects for post-treatment periods $t = T_0 + 1, ..., T_0 + T_1$.

1. Using pretreatment data, run quantile regression of $\{Y_t\}_{t=1}^{T_0}$ on $\{\mathbf{X}_t\}_{t=1}^{T_0}$ via random forest (QRF)[10] at quantiles $\alpha/2$ and $1-\alpha/2$.

2. For each posttreatment period $t = T_0 + 1, ..., T_0 + T_1$, compute conditional quantiles $\widehat{Q}_{Y_t}(\alpha/2|\mathbf{X}_t)$ and $\widehat{Q}_{Y_t}(1-\alpha/2|\mathbf{X}_t)$ using results from Step 1 and posttreatment data for control units $\{\mathbf{X}_t\}_{t=T_0+1}^{T_0+T_1}$.

3. For each posttreatment period $t = T_0 + 1, ..., T_0 + T_1$, compute $(1-\alpha)$ prediction intervals of treatment effects as $\left[Y_t - \widehat{Q}_{Y_t}(1-\alpha/2|\mathbf{X}_t), Y_t - \widehat{Q}_{Y_t}(\alpha/2|\mathbf{X}_t)\right]$.

Remark 3.3. *The estimated quantile function provides an alternative point estimator for the treatment effect. In particular, given the Random Forest estimator of $F(y|\mathbf{X})$, $\widehat{F}(y|\mathbf{X})$, we can obtain the mean prediction based on $\widehat{m}(\mathbf{X}) = \int y d\widehat{F}(y|\mathbf{X})$, and thus the estimator of the treatment effect is given by $\widehat{\Delta}_{1t} = Y_{1t} - \widehat{m}(\mathbf{X}_t)$. Alternatively, if we set $\alpha = 0.5$, then $\widehat{Q}_{Y_t}(\alpha|\mathbf{X}_t)$ corresponds to the median prediction of Y_t conditional on $\mathbf{X}_t$. Based on the median prediction of the counterfactual outcome, we can also construct the median prediction of the treatment effect, which is given by $\widehat{\Delta}_{1t} = Y_{1t} - \widehat{Q}_{Y_t}(0.5|\mathbf{X}_t)$. Compared with the mean prediction of Y_t given in Remark 3.2, the median prediction $\widehat{Q}_{Y_t}(0.5|\mathbf{X}_t)$ is often more robust to outliers of the data set.*

Remark 3.4. *In Athey et al. (2019)'s approach, the gradient of the check function as in Koenker and Bassett (1978) is used to define a pseudo-outcome, which is then used for splitting and growing regression trees. Their gradient tree-based approach requires the trees to be "honest", which is accomplished by splitting the training set into two distinct parts, one part for growing trees only (i.e., estimating the tree structure), and the other part for computing the mean value within each leaf.*

[10]To guarantee the asymptotic validity of the QRF, additional assumptions will be needed when constructing the trees and forests. See Section 3.4 for more details.

3.4 Asymptotic Properties of The QCM

In this section we study the asymptotic properties of the proposed QCM. The validity of the proposed prediction interval depends crucially on the consistency of QRF in the presence of weak dependence and high dimensionality. In the following analysis, for any new test point $\mathbf{X}_0$, we use $\widehat{F}_\theta(y|\mathbf{X}_0) = \sum_{t=1}^{T_0} w_t(\mathbf{X}_0,\theta)\mathbf{1}(Y_t \le y)$ to denote the prediction of $F(y|\mathbf{X}_0)$ based on an individual tree with random feature selection procedure θ, where the weight $w_t(\mathbf{X}_0,\theta)$ is defined in (3.8). Obviously, the prediction from an ensemble of M trees is given by $\widehat{F}^M(y|\mathbf{X}_0) = M^{-1}\sum_{m=1}^{M}\widehat{F}_{\theta_m}(y|\mathbf{X}_0)$, where θ_m is the random feature selection procedure of the m-th tree. We also define the theoretical forest prediction as $\mathbb{E}_\theta\widehat{F}_\theta(y|\mathbf{X}_0) = \int \widehat{F}_\theta(y|\mathbf{X}_0)\, d\Theta(\theta)$, where $\Theta(\theta)$ is the CDF of the random object θ. Essentially, the theoretical forest prediction is the expectation of tree prediction with respect to the random object θ conditional on the data set. Note that when θ_m is iid over m, Law of Large Numbers indicates that $\lim_{M\to\infty} P_\theta\left(\left|\widehat{F}^M(y|\mathbf{X}_0) - \mathbb{E}_\theta\widehat{F}_\theta(y|\mathbf{X}_0)\right| > \varepsilon\right) = 0$ for any $\varepsilon > 0$, where P_θ is the probability with respect to θ. So in this section we will focus on the consistency of $\mathbb{E}_\theta\widehat{F}_\theta(y|\mathbf{X}_0)$.

Throughout the following discussion, the following notations will be frequently used. For any positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, $a_n = o(b_n)$ if $\limsup_n a_n/b_n = 0$ holds, and $a_n = O(b_n)$ if $\limsup_n a_n/b_n < C$ holds for some constant $C \ge 0$. We also write $a_n \sim b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$. For any $\mathbf{X} = (X_1, ..., X_N)^T \in \chi$ and any index set $\mathcal{Q} \subseteq \{1, 2, \cdots, N\}$, define $[\mathbf{X}]_{\mathcal{Q}} = (X_{j_1}, ..., X_{j_{|\mathcal{Q}|}})^T$, where $j_1 < j_2 < \cdots < j_{|\mathcal{Q}|}$ and $j_i \in \mathcal{Q}$ for all i. For any $\mathbf{X}$, $\mathbf{X}'$, and $\mathcal{Q}$, define $\mathbf{V} = [\mathbf{X}]_{\mathcal{Q}} \bigotimes [\mathbf{X}']_{\mathcal{Q}^C}$, where $\mathbf{V} = (V_1, V_2, \cdots, V_N)^T$, and for each i, $V_i = X_i$ if $i \in \mathcal{Q}$ and $V_i = X_i'$ if $i \in \mathcal{Q}^C$. So $\mathbf{X} = [\mathbf{X}]_{\mathcal{Q}} \bigotimes [\mathbf{X}]_{\mathcal{Q}^C}$ holds for all $\mathbf{X}$. For any subset $R \subseteq \chi$, define $[R]_{\mathcal{Q}} = \left\{[\mathbf{X}]_{\mathcal{Q}} : \mathbf{X} \in R\right\}$. If $R = \left\{\mathbf{X} = (X_1, \cdots, X_N) : r_i^- \le X_i \le r_i^+\right\}$, then we write $R = \bigotimes_{i=1}^{N}[r_i^-, r_i^+]$. For an arbitrary set A, define $\operatorname{diam}(A) = \sup_{\mathbf{V},\mathbf{V}'\in A}\|\mathbf{V} - \mathbf{V}'\|$.

To highlight the theoretical findings of this section, we informally state our main theorem as follows:

Theorem. *Suppose that some regularity conditions hold and that $|Y| \le 1$ holds almost surely. Then we have that*

$$\sup_{\mathbf{X}_0\in\mathcal{X}}\sup_{-1\le y\le 1}\left|\mathbb{E}_\theta\widehat{F}_\theta(y|\mathbf{X}_0) - F(y|\mathbf{X}_0)\right| \to_p 0,$$

and

$$\sup_{\mathbf{X}_0\in\mathcal{X}}\sup_{0\le\alpha\le 1}\left|\widehat{Q}_Y(\alpha|\mathbf{X}_0) - Q_Y(\alpha|\mathbf{X}_0)\right| \to_p 0.$$

Readers of interest may refer to the theoretical development of the above theorem in the following

subsections. Basically, the above theorem demonstrates the asymptotic consistency of the QRF under time series setup and high dimensionality, and hence, the asymptotic validity of our QCM method. The proof of the above theorem basically consists of three steps. In the first step, we show that $\mathbb{E}_\theta \widehat{F}_\theta\left(y|\,\mathbf{X}_0\right)$ concentrates around $\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$, where $F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$ is the true cumulative distribution of Y conditional on $\mathbf{X} \in R\left(\mathbf{X}_0,\theta\right)$. In the second step, we show that the difference between $\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$ and our target $F\left(y|\,\mathbf{X}_0\right)$ is upper bounded. Finally, in the third step, we show that such upper bound degenerates to zero as the number of splits increases.

3.4.1 Concentration of Forest Prediction

This subsection is devoted to the derivation of the concentration bounds of the forest prediction. We first make the following technical assumptions.

Assumption 3.1. *The feature space is $\chi = [0,1]^N$. $(Y_t, \mathbf{X}_t)$ is identically distributed over t with joint density $f\left(Y,\mathbf{X}\right)$. Moreover, there exists $\zeta > 1$ such that $\zeta^{-1} < f_{\mathbf{X}}\left(\mathbf{X}\right) < \zeta$, where $f_{\mathbf{X}}\left(\mathbf{X}\right)$ is the marginal density of $\mathbf{X}_t$.*

For convenience, we standardize the feature space χ to $[0,1]^N$. This is a common practice in the literature (Meinshausen, 2006; Biau, 2012; Scornet et al., 2015) since the forest prediction is invariant to any monotone transformation of the feature space. Assumption 3.1 also requires that the marginal density of $\mathbf{X}_t$ is bounded from below and above, a standing example of which is that $\mathbf{X}_t$ is uniformly distributed over χ. In general, it rules out the situation where $\mathbf{X}_t$ has zero probability on a subset of χ with positive Lebesgue measure.

Remark 3.5. *The stationary distribution requirement imposed in Assumption 3.1 is standard in the literature of Random Forest, while it rules out the existence of nonstationarity such as cointegration. As argued in Section 3.3.1, under appropriate assumptions on the nature of nonstationarity, we can transform the data to restore stationarity, and apply our proposed method to the transformed data. Of course, additional assumptions about the model are needed to pursue investigation along this direction. In the current paper, we focus our attension on the stationary data.*

Assumption 3.2. *$\{Y_t, \mathbf{X}_t\}_{t=1}^{\infty}$ is α-mixing with mixing parameters α_t satisfying*

$$\alpha_t \leq C_1 \cdot \rho^{-t}, \tag{3.12}$$

where C_1 is a constant and $\rho > 1$.

Assumption 3.2 requires that $\{Y_t, \mathbf{X}_t\}_{t=1}^{\infty}$ is a strong mixing process with exponentially decaying mixing parameters, which nests many commonly used time series processes such as ARMA(p, q) processes. As shown in Liebscher (1996), the tail behavior of the sum of α-mixing dependent random variables relies on the α-mixing parameters. So this assumption is mainly used to obtain the exponential bound on the tail probability for the sums of random variables. In principle, such an assumption can be weakened to $\alpha_t \leq C \cdot t^{-\beta}$ for some $\beta > 0$, but on this condition we have to require that the dimension of covariate N to increase at a slower rate than that in Assumption 3.4.

Next we make assumptions on the tree structure. Define *splitting level* as the maximum number of nodes that any input has to pass to reach the terminal node. A tree is called *balanced* if the maximum and minimum numbers of nodes that inputs have to pass to reach the terminal nodes are equivalent. We make the following assumptions.

Assumption 3.3. *The tree is balanced. For any parent node* $\bigotimes_{i=1}^{N}[r_i^-, r_i^+]$ *and splitting direction* j*, the splitting point lies within the interval*

$$\left[\left(1-\xi^{-1}\right) r_j^- + \xi^{-1} r_j^+,\ \xi^{-1} r_j^- + \left(1-\xi^{-1}\right) r_j^+\right]$$

with some $\xi > 2$.

Assumption 3.3 imposes two restrictions on the structure of the individual trees. The first requirement of balanced tree can be easily weakened to restrictions on the minimum and maximum cutting numbers for each terminal node. While for convenience, we stick with the balanced-tree assumption in the following development. Assumption 3.3 also requires that any potential splitting point lies in the $\left[\xi^{-1}, 1-\xi^{-1}\right]$ region with $\xi > 2$ for any node and splitting direction, which seems to slightly deviate from the existing literature. In the existing literature, it is usually assumed that each child node contains at least $\widetilde{\xi}^{-1}$ proportion of the total observations in the parent node for some $\widetilde{\xi} > 2$ (Meinshausen, 2006; Wager and Walther, 2016). However, in the Appendix, we show the following proposition[11].

Proposition 3.1. *Suppose that Assumption 3.1, Assumption 3.2, and Assumption 3.4 hold, then: (A) If further, in each round of split, each child node must contain at least* $\widetilde{\xi}^{-1}$ *proportion of the data points in the parent node, where* $\widetilde{\xi} > 2$*, then with probability going to 1, for any parent node* $R = \bigotimes_{i=1}^{N}[r_i^-, r_i^+]$ *and splitting direction* j*, the splitting point lies within the interval*

[11] We thank one anonymous referee for motivating us to think about how our Assumption 3.3 is connected to the existing literature. Such connection is formally stated in Proposition 3.1.

$\left[\left(1-\xi^{-1}\right) r_j^- + \xi^{-1} r_j^+,\ \xi^{-1} r_j^- + \left(1-\xi^{-1}\right) r_j^+\right]$ *where* $\xi > 1.1\zeta^2\widetilde{\xi}$*; (B) If Assumption 3.3 further holds, then with probability going to 1, each child node contains at least* $\widetilde{\xi}^{-1}$ *proportion of data points in the parent node, where* $\widetilde{\xi} > 0.9^{-1}\zeta^2\xi$.

Proposition 3.1 shows that asymptotically, our Assumption 3.3 is equivalent to the conventional assumptions made in the existing literature. In principle, Assumption 3.3 provides a convenient control for the decreasing speed of the volume of the terminal nodes. Note that under Assumption 3.1 and Assumption 3.3, for any node R that may appear in the first k rounds of splits, we have that $\mu(R) \geq \xi^{-k}$, where $\mu(\cdot)$ is the Lebesgue measure, and consequently, $P(R) \geq \zeta^{-1}\xi^{-k}$.

Finally, we assume the following.

Assumption 3.4. *The splitting level* k *and the dimension of covariates* N *satisfy: (A) there exists a constant* $\beta \in \left(0, \frac{1}{3}\right)$ *such that* $N = O(\exp(T_0^{\beta}))$*, and (B)*

$$\pi(k, N, T_0) :\equiv (2\xi)^k \cdot \sqrt{\frac{(\log k + \log N)\log^3 T_0 (\log\log T_0)}{T_0}} \to 0, \tag{3.13}$$

Assumption 3.4(A) restricts the increasing speed of the dimension of the covariates. We allow for ultrahigh-dimensional data set. Assumption 3.4(B) mainly controls the increasing speed of the splitting level k. It requires that $2\log(2\xi)\, k + \log\log N + 3\log\log T_0 + \log\log\log T_0 - \log T_0 \to -\infty$. So the splitting level increases at a speed no faster than $\log T_0 - \log\log N$.

Based on the above assumptions, we have the following result.

Theorem 3.1. *Under Assumption 3.1–Assumption 3.4, there exists a positive constant* C *such that*

$$\lim_{T_0\to\infty} P\left[\sup_{\mathbf{X}_0\in\chi}\sup_{y\in\mathbb{R}}\left|\mathbb{E}_\theta \widehat{F}_\theta(y|\,\mathbf{X}_0) - \mathbb{E}_\theta F(y|\,R(\mathbf{X}_0,\theta))\right| \leq C\cdot\pi(k,N,T_0)\right] = 1. \tag{3.14}$$

Theorem 3.1 indicates that uniformly with respect to the input $\mathbf{X}_0$ and y, the deviation of forest prediction $\mathbb{E}_\theta \widehat{F}_\theta(y|\,\mathbf{X}_0)$ from $\mathbb{E}_\theta F(y|\,R(\mathbf{X}_0,\theta))$ is at most at the rate of $\pi(k,N,T_0)$, which degenerates to zero according to Assumption 3.4. When $N = O(T_0^{\alpha})$ for some $0 < \alpha < \infty$, the convergence rate simplifies to $(2\xi)^k \log^2 T_0 \sqrt{\log\log T_0/T_0}$. Note that given N and T_0, the splitting level k affects the convergence rate $\pi(k,N,T_0)$ exponentially. This is mainly because as k increases, the volume of each terminal leaf decreases exponentially. A smaller node contains fewer observations, based on which the estimation results of the conditional distribution may be more inaccurate. The dimension of

the covariate N affects the convergence rate mainly through complicating the set of terminal leaves. As the number of potential terminal leaves increases, the space over which the supreme operator is performed becomes more complicated, and finally the maximum deviation increases.

We finally compare our results with those in the existing literature. In Wager and Walther (2016), the concentration of forest prediction is of order $\sqrt{\log(T_0/\varsigma)(\log(N\varsigma)+\log\log(T_0))/\varsigma}$, where ς is the minimum leaf size. Since Wager and Walther (2016) assumes that each child node contains at least α proportion of the data points in parent node, there holds $\varsigma \geq \alpha^k T_0$, where k is the number of splits according to our notation. When the equality holds exactly, we have that

$$\begin{aligned}
&\sqrt{\left(\log(T_0/\varsigma)(\log(N\varsigma)+\log\log(T_0))\right)/\varsigma} \\
&= \left(\sqrt{1/\alpha}\right)^k \sqrt{\left(\log(1/\alpha)k(\log N + k\log(\alpha)+\log T_0+\log\log(T_0))\right)/T_0} \\
&\geq C\left(\sqrt{1/\alpha}\right)^k \sqrt{\left(k(\log N+\log T_0)\right)/T_0}.
\end{aligned}$$

for T_0 sufficiently large.

3.4.2 Bounds on the Bias

In the previous subsection, we have shown that the forest prediction $\mathbb{E}_\theta \widehat{F}_\theta\left(y|\,\mathbf{X}_0\right)$ concentrates around $\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$. In principle, the concentration bounds demonstrate how fluctuating our QRF estimator is, while $\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$ is still different from our primary target $F\left(y|\,\mathbf{X}_0\right)$. The task of this subsection is to provide a simple upper bound on the deviation of $\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)$ from $F\left(y|\,\mathbf{X}_0\right)$. In particular, we will work on the following distance

$$\sup_{\mathbf{X}_0\in\mathcal{X}}\sup_{y\in\mathbb{R}}\left|\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)-F\left(y|\,\mathbf{X}_0\right)\right|. \tag{3.15}$$

Before we proceed, we make some further restrictions on the signal structure under high dimensionality.

Assumption 3.5. *There exists an index set $\mathcal{Q}$ such that for any $\mathbf{X}$ and Y, there holds*

$$f_{Y|\mathbf{X}}\left(Y|\mathbf{X}\right)=f_{Y|[\mathbf{X}]_\mathcal{Q}}\left(Y\left|[\mathbf{X}]_\mathcal{Q}\right.\right),$$

where $f_{Y|\mathbf{V}}$ is the density of Y conditional on $\mathbf{V}$. Moreover, for any $\mathbf{X}$, $\mathbf{X}'$, and $\mathbf{X}''$, there holds

$$\left|f\left(Y,[\mathbf{X}]_{\mathcal{Q}}\bigotimes[\mathbf{X}'']_{\mathcal{Q}^C}\right)-f\left(Y,[\mathbf{X}']_{\mathcal{Q}}\bigotimes[\mathbf{X}'']_{\mathcal{Q}^C}\right)\right|\leq L\left(Y\right)\cdot\left\|[\mathbf{X}]_{\mathcal{Q}}-[\mathbf{X}']_{\mathcal{Q}}\right\|, \tag{3.16}$$

with $\int_{\mathbb{R}}L\left(y\right)dy=L<\infty$ for any $Y\in\mathbb{R}$.

Assumption 3.5 first restricts the signal structure of the true data generating process. Under Assumption 3.5, we have

$$f_{[\mathbf{X}]_{\mathcal{Q}^C},Y|[\mathbf{X}]_{\mathcal{Q}}}\left(\left[\mathbf{X}\right]_{\mathcal{Q}^C},Y|\left[\mathbf{X}\right]_{\mathcal{Q}}\right)=f_{[\mathbf{X}]_{\mathcal{Q}^C}|[\mathbf{X}]_{\mathcal{Q}}}\left(\left[\mathbf{X}\right]_{\mathcal{Q}^C}|\left[\mathbf{X}\right]_{\mathcal{Q}}\right)f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y|\left[\mathbf{X}\right]_{\mathcal{Q}}\right).$$

So Y is independent of $[\mathbf{X}]_{\mathcal{Q}^C}$ conditional on $[\mathbf{X}]_{\mathcal{Q}}$. When the index set $\mathcal{Q}$ is fixed while the dimension of $\mathbf{X}$ is allowed to increase with sample size T_0, such an assumption can be interpreted as a sparsity condition, which is generally used to deal with high-dimensional case. Assumption 3.5 also requires that for any fixed Y, the joint density of $(Y,\mathbf{X})$ is $L\left(Y\right)$-Lipschitz with respect to $[\mathbf{X}]_{\mathcal{Q}}$. Define $f_{\mathcal{Q}}\left(Y,[\mathbf{X}]_{\mathcal{Q}}\right)$ as the joint density of $\left(Y,[\mathbf{X}]_{\mathcal{Q}}\right)$. Under Assumption 3.5, we have $\left|f_{\mathcal{Q}}\left(Y,[\mathbf{X}]_{\mathcal{Q}}\right)-f_{\mathcal{Q}}\left(Y,[\mathbf{X}']_{\mathcal{Q}}\right)\right|\leq L\left(Y\right)\cdot\left\|[\mathbf{X}]_{\mathcal{Q}}-[\mathbf{X}']_{\mathcal{Q}}\right\|$, so $f_{\mathcal{Q}}\left(Y,[\mathbf{X}]_{\mathcal{Q}}\right)$ is also $L\left(Y\right)$-Lipschitz with respect to $[\mathbf{X}]_{\mathcal{Q}}$. Moreover, define $f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)$ as the marginal density of $[\mathbf{X}]_{\mathcal{Q}}$. There holds $\left|f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)-f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}']_{\mathcal{Q}}\right)\right|\leq L\cdot\left\|[\mathbf{X}]_{\mathcal{Q}}-[\mathbf{X}']_{\mathcal{Q}}\right\|$, which implies that $f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)$ is L-Lipschitz.

Based on Assumption 3.5, we have the following result.

Theorem 3.2. *Under Assumption 3.1 and Assumption 3.5, there exists a constant C such that*

$$\sup_{\mathbf{X}_0\in\mathcal{X}}\sup_{y\in\mathbb{R}}\left|\mathbb{E}_\theta F\left(y|\,R\left(\mathbf{X}_0,\theta\right)\right)-F\left(y|\,\mathbf{X}_0\right)\right|\leq C\cdot\sup_{\mathbf{X}_0\in\mathcal{X}}\mathbb{E}_\theta\mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right).$$

Theorem 3.2 implies that based on the sparsity assumption, we can construct an upper bound for the bias of the forest prediction that depends only on the signal variables. In the following subsection, we will show that such upper bound degenerates to 0 under some further conditions, and hence prove the consistency of the QRF.

3.4.3 Consistency

In the previous subsection we have shown that the bias of the QRF is, up to a constant term, bounded by $\sup_{\mathbf{X}_0\in\mathcal{X}}\mathbb{E}_\theta\mathrm{diam}([R\left(\mathbf{X}_0,\theta\right)]_{\mathcal{Q}})$. When the number of the covariates is fixed and each covariate is

split with a probability bounded away from 0 (e.g. Meinshausen, 2006), $\sup_{\mathbf{X}_0 \in \mathcal{X}} \mathbb{E}_\theta \text{diam}([R(\mathbf{X}_0, \theta)]_\mathcal{Q}) \to 0$ as $k \to \infty$ naturally holds under Assumption 3.3. While in the high-dimensional scenario where $N \gg k$, if we do not distinguish between the signal variables and noise variables (for example, pick each covariate with the same probability), we may end up making too many splits along the noise variables, which does not effectively decrease the bias.

Many attempts have been made to avoid the above problem. For example, in Wager and Walther (2016), a Guess-and-Check tree structure is imposed to screen out noise variables that are independent of the response. However, such framework does not directly apply to our setup. For the panel data driven by common factors, noise variables are not necessarily independent of the response. On this condition, when a noise variable is picked, splitting on it will lead to significant child nodes' differences, hence the noise variable will be unblocked and we may end up making too many cuts along the noise variables.

Going back to the tree splitting algorithm discussed in Section 3.3.2, we now consider a specific test input $\mathbf{X}_0$ and one individual tree with random feature selection vector θ. When splitting a parent node that contains $\mathbf{X}_0$, m_{try} features are randomly selected based on the realization of θ and then compared. For some fixed positive integer d, suppose that $m_{try} \sim N$ and we conduct k rounds of splits. Then the probability that the j-th covariate is selected as a candidate for split with less than or equal to d times goes to 0 as $k \to \infty$. This implies that as the split proceeds, the number of rounds in which a particular signal variable is taken as the potential splitting direction will increase to infinity with probability going to 1 (probability with respect to θ). So the remaining task is to investigate whether each signal variable is indeed selected and split with growing number of times. For simplicity, we assume that $|Y| \leq 1$ throughout this section. Such upper bound can be replaced with any positive constant. The upper boundedness of the response is for technical proofs and is mainly used to obtain exponential concentration inequalities for dependent data. Similar boundedness assumption is also imposed in Wager and Walther (2016). It is possible to replace such condition by sub-Gaussianity of Y_t.

As was discussed in Section 3.3.2, given any rectangle $R = \bigotimes_{i=1}^{N} \left[r_i^-, r_i^+\right] \subseteq \chi$, the splitting criterion is to maximize $\mathcal{I}\widehat{(R, i, \lambda)}$ with respect to i and λ. Define the population counterpart of $\mathcal{I}\widehat{(R, i, \lambda)}$ as

follows

$$\mathcal{I}(R,i,\lambda) = \sigma^2(Y|R) - P(R\cap\{X_i < r_{i,\lambda}\}|R)\,\sigma^2(Y|R\cap\{X_i < r_{i,\lambda}\}) \\ - P(R\cap\{X_i \geq r_{i,\lambda}\}|R)\,\sigma^2(Y|R\cap\{X_i \geq r_{i,\lambda}\}). \tag{3.17}$$

(3.17) is obviously the splitting criterion we would maximize if we were able to observe the whole population. According to Lemma 3.7 in the Appendix, the empirical criterion (3.6) constitutes a good approximation of (3.17) uniformly with respect to rectangles R, splitting direction i, and splitting location λ. Motivated by such result, to guarantee that each signal variable can be selected and cut for increasing number of times, we only need to impose more restrictions on $\mathcal{I}(R,i,\lambda)$. This is done by the following Assumption 3.6 and Assumption 3.7.

Assumption 3.6. *$\mathcal{Q}$ is fixed. Define*

$$\delta(\eta) = \inf_{i\in\mathcal{Q}} \inf_{\left\{R=\bigotimes_{j=1}^N [r_i^-, r_i^+]\subseteq\chi : r_i^+ - r_i^- = \eta\right\}} \mathcal{I}(R,i,1/2), \tag{3.18}$$

there holds $\delta(\eta) > 0$ for any $\eta > 0$.

Assumption 3.6 imposes restrictions on the lower bound of $\mathcal{I}(R,i,\lambda)$ for signal variables. It requires that the total variation of Y contributed by the variation of signal X_i can not be fully explained by the combination of the remaining covariates, and the unexplained variation is uniformly lower-bounded by a positive function depending only on the length of the interval where X_i takes value. So even after we have split along the same signal variable for sufficiently many times, we will still find significantly different child nodes if we continue to cut along such direction.

Remark 3.6. *We provide two examples for Assumption 3.6. Consider the following linear model $Y = \sum_{i\in\mathcal{Q}} X_i + \varepsilon$, where X_i's and ε are mutually independent. For any $R = \bigotimes_{j=1}^N [r_j^-, r_j^+]$, we have that $\sigma^2(Y|R) = \sum_{i\in\mathcal{Q}} \sigma^2\left(X_i | X_i \in \left[r_i^-, r_i^+\right]\right) + \sigma_\varepsilon^2$. Suppose further that $X_i \sim U(0,1)$, we have $\mathcal{I}(R,i,1/2) = \left(r_i^+ - r_i^-\right)^2/16$. Consider another example $Y = \prod_{i\in\mathcal{Q}}(1+X_i) + \varepsilon$, where X_i's and ε are mutually independent, and $X_i \sim U(0,1)$. For any $R = \bigotimes_{j=1}^N [r_j^-, r_j^+]$, we have that $\mathcal{I}(R,i,1/2) = (\prod_{j\in\mathcal{Q},j\neq i}(1+(r_j^+ + r_j^-)/2))((1+r_i^+/4+3r_i^-/4)^2/2 + (1+3r_i^+/4+r_i^-/4)^2/2 - (1+r_i^+/2+r_i^-/2)^2)$. Then due to the fact that $(1+x_1)^2/2 + (1+x_2)^2/2 - (1+(x_1+x_2)/2)^2 = (x_1+x_2)^2/4$, we have that $\mathcal{I}(R,i,1/2) \geq (\prod_{j\in\mathcal{Q},j\neq i}(1+(r_j^+ + r_j^-)/2)) \cdot (r_i^+ - r_i^-)^2/16 \geq (r_i^+ - r_i^-)^2/16$.*

We finally make another assumption on the information of the noise variables.

Assumption 3.7. *If N is diverging, there exists a fixed set $\mathcal{Q}_1 \supseteq \mathcal{Q}$ such that for any $R \subseteq \chi$ and $i \notin \mathcal{Q}_1$, there holds $\sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, i, \lambda) < \omega \max_{j \in \mathcal{Q}_1} \sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j, \lambda)$, where $0 < \omega < 1$. Let $\mathcal{Q}_1 = \{1, 2, \cdots, N\}$ when N is fixed. For N either fixed or diverging, the Lipschitz condition (3.16) also holds for $\mathcal{Q}_1$*

Assumption 3.7 implies that, when there is increasing number of covariates, although splitting on a noise variable can lead to larger differences in the child nodes compared with splitting on any of the signal variables, there are not infinitely many such covariates. Note that such an assumption does not rule out the possibility that all the noise variables are correlated with the response; instead, we only require that the number of "more informative" noise variables is finite. On this condition, if we had observed the whole population, we would never split on the covariates outside $\mathcal{Q}_1$ if we can choose from all the variables in $\mathcal{Q}_1$.

Remark 3.7. *Assumption 3.7 can be regarded as an extension of the scenario where all the noise variables are independent of the response. In fact if so, we have $\sigma^2(Y|R \cap \{X_i < r_{i,\lambda}\}) = \sigma^2(Y|R \cap \{X_i \geq r_{i,\lambda}\}) = \sigma^2(Y|R)$ for any R and $i \notin \mathcal{Q}$, so $\mathcal{I}(R, i, \lambda) = 0$ for any $\lambda \in [0, 1]$. If Assumption 3.6 further holds, we have $0 = \mathcal{I}(R, i, \lambda) < \omega \mathcal{I}(R, j, 1/2) \leq \sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j, \lambda)$ for any $j \in \mathcal{Q}$.*

Based on the restrictions on the population splitting criterion as were made in Assumption 3.6 and Assumption 3.7, we study the behavior of the empirical splits and the bias of QRF. We first provide two lemmas that are useful when N is diverging. Define $\Psi(\mathbf{X}_0, k, d)$ as the collection of θ such that in the first k rounds of splits along the nodes that contain $\mathbf{X}_0$, there are at least d rounds in which all the covariates in $\mathcal{Q}_1$ are simultaneously selected as candidates. Note that if we we choose $m_{try} \sim N$, since $N \to \infty$, $m_{try} \gg |\mathcal{Q}_1|$ for N sufficiently large and hence $\lim_{k \to \infty} P_\theta[\Psi(\mathbf{X}_0, k, d)] = 1$ for any fixed d. Moreover, given the feature selection procedure θ and the total number of splits k, define $\mathcal{N}(\mathbf{X}_0, \theta, k, j)$ as the number of cuts over covariate j, which is a random variable. We have the following result.

Lemma 3.1. *Suppose Assumption 3.1–Assumption 3.7 hold. Suppose moreover $|Y| \leq 1$ holds almost surely, N is diverging, and $m_{try} \sim N$. For any fixed k and $d \leq k$, there holds*

$$\lim_{T_0 \to \infty} P\left[\inf_{\mathbf{X}_0 \in \mathcal{X}: \Psi(\mathbf{X}_0, k, d) \neq \emptyset} \inf_{\theta \in \Psi(\mathbf{X}_0 k, d)} \sum_{j \in \mathcal{Q}_1} \mathcal{N}(\mathbf{X}_0, \theta, k, j) \geq d\right] = 1.$$

The idea of Lemma 3.1 is intuitively discussed as follows. Under Assumption 3.7, when all covariates

in $\mathcal{Q}_1$ are selected, we will choose one covariate in $\mathcal{Q}_1$ to split if we can observe (3.17). So if $\mathcal{Q}_1$ is selected d times, there must be at least d rounds in which splits take place in $\mathcal{Q}_1$. Although in practice we can not observe (3.17), its empirical counterpart (3.6) constitutes a good estimate for (3.17), so the split based on (3.6) will be asymptotically identical to that based on (3.17).

In the Appendix, we also show that $\mathcal{I}(R, i, \lambda)$ with $i \in \mathcal{Q}_1$ decreases exponentially uniformly with respect to R and λ as the number of splits along such covariate increases. Combine such result with the lower bound on $\mathcal{I}(R, i, \lambda)$ as assumed in Assumption 3.6, it is then intuitive to expect that as long as d is sufficiently large, each covariate in $\mathcal{Q}$ will be cut with no less than a predetermined number of times. Such a result is formally stated in the following lemma.

Lemma 3.2. *Suppose Assumption 3.1-Assumption 3.7 hold. Suppose moreover $|Y| \leq 1$ holds almost surely, N is diverging, and $m_{try} \sim N$. For any fixed $d \in \mathbb{N}$, let d^* satisfy $2C \cdot \left(1 - \xi^{-1}\right)^{d^*-1} < \inf_{\eta \geq \xi^{-(d-1)}} \delta(\eta)$, where $C = C(d^*)$ is a constant depending on d^*. For any fixed k, there holds*

$$\lim_{T_0 \to \infty} P\left[\inf_{\mathbf{X}_0 \in \mathcal{X}: \Psi(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\}) \neq \emptyset} \inf_{\theta \in \Psi(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\})} \min_{j \in \mathcal{Q}} \mathcal{N}(\theta, k, j) \geq d\right] = 1.$$

Based on Lemma 3.1 and Lemma 3.2, we have the following result which is crucial in showing the consistency of QRF.

Lemma 3.3. *Suppose Assumption 3.1–Assumption 3.7 hold. Suppose moreover $|Y| \leq 1$ holds almost surely, $m_{try} \sim N$ and $k \to \infty$. Then for any positive integer $d > 0$, we have*

$$p\lim_{T_0 \to \infty} \inf_{\mathbf{X}_0 \in \mathcal{X}} P_\theta\left[\min_{j \in \mathcal{Q}} \mathcal{N}(\mathbf{X}_0, \theta, k, j) \geq d\right] = 1,$$

where P_θ is the probability measure with respect to θ.

Lemma 3.3 is the key result of this section. It demonstrates that under the QRF algorithm, no matter whether the number of covaraites N is diverging or fixed, as the number of split k increases the average (with respect to θ) number of splits over each signal variable will increase to exceed any fixed integer with probability going to 1. Such a result implies that with probability going to 1, the average number of splits over signals will go to infinity, and consequently, the bias also degenerates to 0 according to Theorem 3.2.

Based on Lemma 3.3, now we can demonstrate the consistency of the QRF.

Theorem 3.3. *Suppose Assumption 3.1–Assumption 3.7 hold. Moreover, suppose $|Y| \leq 1$ holds*

almost surely, $m_{try} \sim N$ *and* $k \to \infty$*. Then we have that*

$$\sup_{\mathbf{X}_0 \in \mathcal{X}} \sup_{-1 \leq y \leq 1} \left| \mathbb{E}_\theta \widehat{F}_\theta \left(y | \mathbf{X}_0 \right) - F \left(y | \mathbf{X}_0 \right) \right| \to_p 0.$$

Suppose further that $\sup_{\mathbf{X}_0 \in \mathcal{X}, y} F_y \left(y | \mathbf{X}_0 \right) < \infty$*,* $\inf_{\mathbf{X}_0 \in \mathcal{X}, |y| \leq c} F_y \left(y | \mathbf{X}_0 \right) > 0$ *for all* $0 < c < 1$ *and* $\left(\inf_{\mathbf{X}_0 \in \mathcal{X}, |y| \leq c} F_y \left(y | \mathbf{X}_0 \right) \right)^{-1} \cdot (1 - c) \cdot \sup_{\mathbf{X}_0 \in \mathcal{X}, |y| \geq c} F_y \left(y | \mathbf{X}_0 \right) \to 0$ *for* $c \to 1$*, where* $F_y(y|\mathbf{X}_0)$ *is the partial derivative of* $F(y|\mathbf{X}_0)$ *with respect to* y*. Then*

$$\sup_{\mathbf{X}_0 \in \mathcal{X}} \sup_{0 \leq \alpha \leq 1} \left| \widehat{Q}_Y \left(\alpha | \mathbf{X}_0 \right) - Q_Y \left(\alpha | \mathbf{X}_0 \right) \right| \to_p 0.$$

Remark 3.8. *(A) Our proof technique can be easily extended to prove the point consistency of the Random Forest prediction of the conditional mean in the high-dimensional scenario. So under the similar conditions, we can show the consistency of the treatment effect estimator discussed in Remark 3.2. Such a result is not trivial compared with Wager and Walther (2016) since we do not impose Guess-and-Check strcture. (B) The high-level assumptions we use throughout the proof are Assumption 3.6 and Assumption 3.7. In general, both assumptions are difficult to break down to more primitive assumptions, while it is easy to verify them when given specific setups of the data generating processes. Note that when* $\eta = 1$*, Assumption 3.6 is equivalent to the Assumption 4 (monotone signal) in Wager and Walther (2016). (C) Note that the above asymptotic consistency results hold for any fixed index set* $\mathcal{Q}$*. It's also not difficult to see that there exists a slowly increasing sequence of index sets* $\{\mathcal{Q}_{T_0}\}_{T_0=1}^{\infty}$ *such that* $\lim_{T_0 \to \infty} |\mathcal{Q}_{T_0}| = \infty$*, and that the consistency results still hold.*

3.5 Simulations

In this section, we investigate the finite sample performance of the proposed method via Monte Carlo simulations. We consider a wide-range of data-generating processes, including those in Hsiao et al. (2012), then add heteroskedasticity, cross-sectional heteroskedasticity, autocorrelation or within-panel autocorrelation, and nonlinear transformations, followed by DGPs free of any factor structure with or without sparsity. In particualr, following Cattaneo et al. (2021), Chernozhukov et al. (2021b) and Hsiao et al. (2012), we consider the following 13 different data generating processes:

(1) DGP1 (sparse weights). We model the potential outcome without treatment for the first treated

unit as a sparse linear function of the other control units:

$$y_{1t}^0 = \alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0 + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$, and

$$(w_2, ..., w_{N+1}) = (0.2, 0.2, 0.2, 0.2, 0.2, 0, ..., 0).$$

(2) DGP2 (even weights). As a variation of DGP1, consider DGP2 with small effects for all cross-sectional units:

$$y_{1t}^0 = \alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0 + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$, and $(w_2, ..., w_{N+1}) = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$.

(3) DGP3 (single iid factor). The third DGP has a simple i.i.d. factor ($r = 1$):

$$y_{it}^0 = \alpha_i + b_i f_t + u_{it},\ f_t \sim_{iid} N(0,1),$$

where α_i, u_{it}, and f_t are random draws from $N(0,1)$, while factor loadings $b_i's$ are random draws from $N(1,1)$.

(4) DGP4 (two stationary factors). The fourth DGP consists of two ($r = 2$) stationary factors:

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + u_{it},$$

with

$$f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t},\ f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t},$$

where α_i, ε_{1t} and ε_{2t} are random draws from $N(0,1)$, factor loadings b_{i1} and b_{i2} are random draws from $N(1,1)$, while we let $u_{it} = \chi^2(1) - 1$ as a variation.

(5) DGP5 (DGP1 + sine). DGP5 applies a nonlinear sine transformation of DGP1:

$$y_{1t}^0 = \sin\left(\alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0\right) + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$,

$$(w_2, ..., w_{N+1}) = (0.2, 0.2, 0.2, 0.2, 0.2, 0, ..., 0),$$

and $\sin(\cdot)$ is the sine function.

(6) DGP6 (DGP1 + cube). DGP6 applies an unbounded cubic transformation of DGP1:

$$y_{1t}^0 = \left(\alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0\right)^3 + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$, and

$$(w_2, ..., w_{N+1}) = (0.2, 0.2, 0.2, 0.2, 0.2, 0, ..., 0).$$

(7) DGP7 (DGP2 + sine). DGP7 applies a nonlinear sine transformation of DGP2:

$$y_{1t}^0 = \sin\left(\alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0\right) + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$, and $(w_2, ..., w_{N+1}) = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$.

(8) DGP8 (DGP2 + cube). DGP8 applies an unbounded cubic transformation of DGP2:

$$y_{1t}^0 = \left(\alpha_i + \sum_{j=2}^{N+1} w_j y_{jt}^0\right)^3 + u_t,$$

where α_i, y_{jt}^0 and u_t are iid random draws from $N(0,1)$, and $(w_2, ..., w_{N+1}) = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$.

(9) DGP9 (cubic transformation with sparse weights). DGP9 considers the following nonlinear data generating process:

$$y_{1t}^0 = \left(\alpha_1 + \sum_{j=2}^{N+1} w_j (y_{jt}^0)^3 + u_t \right)^{1/3},$$

where α_1, y_{jt}^0 and u_t are $i.i.d.$random draws from $N(0,1)$, and

$$(w_2, ..., w_{N+1}) = (0.2, 0.2, 0.2, 0.2, 0.2, 0, ..., 0).$$

(10) DGP10 (cubic transformation with equal weights). DGP10 considers the following nonlinear data generating process:

$$y_{1t}^0 = \left(\alpha_1 + \sum_{j=2}^{N+1} w_j (y_{jt}^0)^3 + u_t \right)^{1/3},$$

where α_1, y_{jt}^0 and u_t are $i.i.d.$random draws from $N(0,1)$, and $(w_2, ..., w_{N+1}) = (\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$.

(11) DGP11 (DGP4 + heteroskedasticity). DGP11 is a variation of DGP4 with cross-sectional heteroskedasticity, but without within-panel autocorrelation:

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + 0.2 h_i u_{it},$$

with

$$f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t}, \;\; f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t},$$

where $h_i's$ are random draws from uniform distribution $U(0,10)$, factor loadings b_{i1} and b_{i2} are random draws from $N(1,1)$, while α_i, u_{it}, ε_{1t} and ε_{2t} are random draws from $N(0,1)$.

(12) DGP12 (DGP4 + autocorrelation). DGP12 is a varation of DGP4 with within-panel autocorrelation, but no cross-sectional heteroskedasticity:

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + u_{it},\ u_{it} = \rho_i u_{i,t-1} + v_{it}$$

with

$$f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t},\ f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t},$$

where v_{it} are random draws from $N(0,1)$, while $\rho_i's$ are random draws from uniform distribution $U(0,1)$. Note that u_{it} is standardized to have unit variance in order to maintain homoskedasticity across units.

(13) DGP13 (DGP4 + HAC). DGP13 is a variation of DGP4 with both cross-sectional heteroskedasticity and within-panel autocorrelation:

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + 0.2 h_i u_{it},\ u_{it} = \rho_i u_{i,t-1} + v_{it}$$
$$f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t},\ f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t},$$

where $h_i's$ are generated as in DGP9, while v_{it} and ρ_i are generated as in DGP10, and u_{it} is again standardized to have a unit variance.

In all simulations, the true treatment effect of size 1 is assumed to impact the first unit after the treatment, while all other units receive no treatment. The nominal coverage rate is 95%. The coverage probability (i.e., empirical coverage) is computed as the frequency that the constructed prediction intervals contain the true treatment effect in 1000 simulations. When using the R package quantregForest for the implementation of quantile random forest, we set the number of trees to be 1000, the minimum size of terminal nodes to be 10, and turn of the bootstrap part of the algorithm[12].

Table 3.1 reports the performance of the QCM approach by varying the number of pretreatment periods from 10 through 90, with $N = 30$. To save space, additional results with different number of control units N are reported in the Supplementary Material[13]. The results are quite similar. The results reported in Table 3.1 demonstrate great finite-sample properties of QCM prediction intervals across different DGPs. The coverage probabilities reach around 0.9 even with $T_0 = 20$. Overall, the coverage probabilities appear to approach the nominal coverage rate of 0.95 as T_0 becomes

[12]We find that the minimum size of terminal nodes of 10 is slightly better than the default setting of 5 for our data structure.

[13]The readers are also referred to an earlier version of this paper for more results on the Monte Carlo experiments.

Table 3.1: Coverage Probabilities of QCM ($N = 30$)

DGP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
1	0.822	0.908	0.921	0.916	0.921	0.929	0.918	0.935	0.929
2	0.829	0.888	0.927	0.911	0.91	0.925	0.947	0.942	0.925
3	0.799	0.889	0.92	0.917	0.943	0.943	0.933	0.947	0.943
4	0.781	0.877	0.911	0.919	0.934	0.947	0.953	0.948	0.956
5	0.814	0.895	0.918	0.912	0.903	0.927	0.921	0.917	0.934
6	0.838	0.892	0.908	0.925	0.937	0.923	0.944	0.945	0.932
7	0.821	0.909	0.904	0.917	0.925	0.92	0.927	0.922	0.935
8	0.808	0.883	0.923	0.925	0.921	0.933	0.947	0.94	0.936
9	0.826	0.887	0.911	0.923	0.922	0.917	0.945	0.92	0.927
10	0.816	0.892	0.926	0.929	0.924	0.928	0.928	0.916	0.919
11	0.798	0.883	0.922	0.925	0.934	0.945	0.955	0.96	0.96
12	0.778	0.879	0.928	0.914	0.937	0.948	0.96	0.937	0.953
13	0.786	0.865	0.909	0.93	0.938	0.941	0.952	0.96	0.943

Note: T_0 is the number of pretreatment periods. The nominal coverage rate is 95%.

large. These simulations confirm that QCM prediction intervals have excellent coverage properties even in small samples with the number of pretreatment periods as small as 30. Moreover, QCM prediction intervals are shown to be robust to heteroskedasticity, autocorrelation, sparsity and model misspecification, since quantile random forest is nonparametric by nature as an ensemble learning based on decision trees.

Next, we compare the performance of the prediction intervals derived from the quantile control method (QCM) with existing approaches in the literature. In particular, we compare the finite sample performance of the proposed procedure with those of Cattaneo et al. (2021) and Chernozhukov et al. (2021b). Cattaneo et al. (2021) propose three approaches to measure the out-of-sample uncertainty of synthetic control methods: the sub-Gaussian approach using concentration inequalities under subgaussian distribution, the location-scale approach relying on location-scale model, and the quantile regresssion approach applying linear quantile regression to the residuals. [14] Chernozhukov et al. (2021b) adopt a conformal inference relying on moving-block permutation of the estimated residuals to hypothesis testing for synthetic control methods, and builds prediction intervals indirectly by test inversion. For the three approaches by Cattaneo et al. (2021), we use the R package `scpi` provided by the authors for implementation. For the implementation of conformal approach by Chernozhukov et al. (2021b), we use the R package `scinference` provided by the authors.

In addition to the above methods, two other methods based on stronger assumptions are also considered in our Monte Carlo. In particular, Fujiki and Hsiao (2015) proposes confidence intervals for regression control method (RCM) under the assumptions of linear factor model, iid disturbances

[14] As commented in Cattaneo et al. (2021), one may also consider using nonparametric quantile regression.

and normality. Bai and Ng (2021) proposes a Tall-Wide algorithm to impute counterfactual outcomes in a similar panel data setting, as well as confidence intervals relying on assumptions of linear factor model, identical distributions (homoskedasticity) and normality. We also include these two procedures for the comparison of the QCM with existing methods.

The sample size is $T_0 = 30$, and the number of control unitsn is $N = 30$. We choose these sizes because they are close to those in empirical applications in the literature, for example, the pre-treatment periods for the California tabacco control study (Abadie et al., 2010) and the Germany reunification study (Abadie et al., 2015) are only 19 and 30 respectively. The simulation results are presented in Table 3.2 and Figure 3.1. Under a small-sample setting with $N = 30$ and $T_0 = 30$, the coverage probabilities of QCM are all above 0.9, and not far from the nominal coverage rate of 0.95. In comparison, prediction intervals of the Gaussian, Location-Scale and Quantile Regression approaches generally undercover, with coverage probabilities being generally below 0.9, sometimes dipping down below 0.8. Moreover, the performance of conformal approach seems unstable, with coverage probabilities fluctuating between severe overcover (e.g., CP reaching 0.999 in DGP7) and severe undercover (e.g., CP dropping to 0.709 in DGP3), despite great performance in DGP8 and DGP12. The RCM-based method performs poorly when the sample size T_0 is small, and undercover in general. The Tall-Wide algorithm based approach has reasonable performance for some simple DGPs, but generally undercover in complicate models. As shown in Figure 3.1, the prediction intervals of QCM clearly dominate other approaches in terms of coverage probabilities.

We finally compare large-sample properties of different approaches by considering the case with $T_0 = 100$ and $N = 30$. While $T_0 = 100$ is unrealistically large in practice, it is helpful for investigating the consistency of prediction intervals. The results are reported in Table 3.3. With a larger T_0, the performance of Gaussian, Location-Scale and Quantile Regression approaches have all improved. In particular, the performance of Location-Scale and Quantile Regression approaches are now comparable to QCM, and even slightly outperform the performance of QCM under simple linear or linear index setups (for example, DGPs 1, 2, 5, 7, and 8). However, these methods are generally dominated by the QCM when the data generating process displays severe nonlinearity (for example, DGPs 9, 10, 11, and 12). Both the RCM-based method and the Tall-Wide algorithm based approach improve when T_0 is large, but still undercover when nonlinearity increases. Such results highlight the robustness of our proposed method. Finally, we note that the coverage probabilities of conformal approach remain unstable and are far away from the nominal rate of 0.95 in many cases.

Table 3.2: Comparing Prediction Intervals ($N = 30$, $T = 30$)

DGP	CP-Q	AL-Q	CP-G	AL-G	CP-LS	AL-LS	CP-QR	AL-QR	CP-C	AL-C	CP-R	AL-R	CP-TW	AL-TW
1	0.909	4.088	0.804	4.968	0.863	4.784	0.887	4.893	0.998	5.065	0.654	3.381	0.905	4.037
2	0.92	6.856	0.87	5.333	0.906	4.581	0.92	5.424	0.778	4.404	0.668	3.321	0.911	4.010
3	0.917	6.822	0.838	0.988	0.877	1.049	0.898	1.216	0.709	1.021	0.662	3.400	0.653	3.813
4	0.914	3.777	0.808	2.345	0.859	3.521	0.884	3.604	0.998	6.186	0.690	4.855	0.649	5.067
5	0.928	5.080	0.832	2.158	0.883	3.172	0.899	3.387	0.867	5.005	0.672	3.348	0.901	3.947
6	0.923	7.916	0.803	5.011	0.874	5.635	0.894	7.407	0.912	9.209	0.673	4.638	0.868	3.921
7	0.922	3.915	0.79	3.315	0.856	3.239	0.877	3.749	0.999	5.966	0.672	3.319	0.904	3.939
8	0.933	8.261	0.788	2.627	0.853	2.927	0.887	1.716	0.942	7.568	0.655	3.660	0.88	3.892
9	0.921	2.467	0.835	6.936	0.857	6.977	0.857	6.118	0.997	11.992	0.705	2.996	0.749	11.256
10	0.907	1.958	0.81	3.504	0.847	3.951	0.854	3.993	0.99	4.144	0.757	2.721	0.777	11.569
11	0.906	3.782	0.763	1.253	0.833	2.163	0.855	1.313	0.998	7.347	0.701	3.499	0.59	3.814
12	0.914	4.779	0.788	3.802	0.855	4.546	0.869	5.604	0.954	11.351	0.607	3.122	0.602	3.265
13	0.921	6.787	0.855	2.901	0.888	3.288	0.914	4.128	0.74	2.362	0.627	2.989	0.587	3.132

Note: The nominal coverage rate is 95%. Q refers to QCM, G refers to Gaussian, LS refers to Location-Scale, QR refers to Quantile Regression, C refers to Conformal, RCM refers to regression control method, and TW refers to Tall-Wide. CP-Q, CP-Gaussian, CP-LScale, CP-QR, CP-C, CP-R, and CP-TW refer to coverage probability of the above methods. AL-Q, AL-Gaussian, AL-LScale, AL-QR, AL-Conform, AL-R, and AL-TW refer to the average length of preditions intervals for these approaches. The above also applies to Table 3.3.

Table 3.3: Comparing Prediction Intervals ($N = 30$, $T = 100$)

DGP	CP-Q	AL-Q	CP-G	AL-G	CP-LS	AL-LS	CP-QR	AL-QR	CP-C	AL-C	CP-R	AL-R	CP-TW	AL-TW
1	0.915	4.114	0.894	3.308	0.957	4.025	0.959	4.182	0.999	4.424	0.914	3.993	0.966	4.348
2	0.934	3.8813	0.861	2.710	0.95	3.391	0.952	3.667	1	4.344	0.912	3.896	0.964	4.251
3	0.943	4.506	0.895	3.785	0.955	4.106	0.963	4.490	0.777	4.545	0.917	4.005	0.948	4.845
4	0.955	6.962	0.909	3.540	0.959	8.660	0.961	10.131	0.834	6.326	0.914	5.690	0.935	6.948
5	0.926	4.003	0.871	2.523	0.95	3.812	0.951	3.840	1	5.325	0.915	3.941	0.961	4.284
6	0.945	8.034	0.891	2.656	0.957	3.230	0.959	3.384	0.944	4.665	0.929	5.238	0.961	4.211
7	0.919	3.815	0.84	2.790	0.927	4.107	0.936	3.949	0.999	4.585	0.914	3.878	0.961	4.246
8	0.918	4.791	0.856	3.401	0.942	3.783	0.951	3.836	0.968	4.304	0.913	4.263	0.959	4.124
9	0.938	2.518	0.989	6.585	0.993	6.624	0.985	5.074	0.962	5.305	0.943	2.566	0.956	13.687
10	0.933	1.955	0.98	3.457	0.985	3.420	0.977	3.412	0.96	4.024	0.941	2.285	0.956	13.705
11	0.964	5.709	0.928	3.110	0.975	3.273	0.975	3.239	0.658	2.803	0.924	3.777	0.873	6.871
12	0.962	5.599	0.96	3.735	0.987	4.768	0.984	4.282	0.683	3.504	0.883	3.76	0.865	4.639
13	0.952	5.573	0.922	6.301	0.975	7.575	0.971	8.071	0.675	5.726	0.893	3.731	0.83	5.597

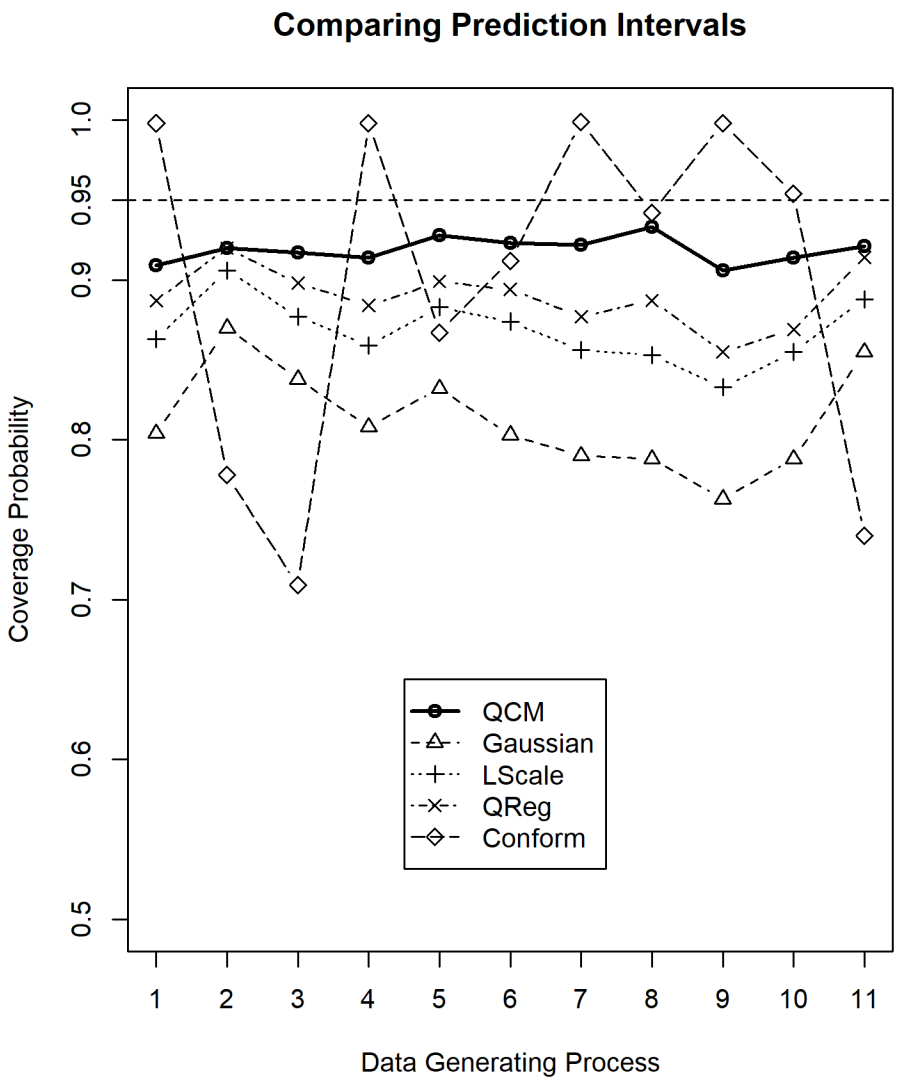


Figure 3.1: Comparing Prediction Intervals

3.6 Empirical Application

In this section, we study the impact on Hong Kong's real GDP growth rate with the implementation of CEPA between mainland China and Hong Kong using quarterly data from 1993Q1 to 2008Q1 for 25 countries and regions including Hong Kong (Hsiao et al., 2012). In June 2003, Hong Kong signed the Closer Economic Partnership Arrangement (CEPA) with mainland China, which went into effect on January 1st, 2004. CEPA aimed to strengthen the linkage between mainland China and Hong Kong by liberalizing trade in services, enhancing cooperation in the area of finance, promoting trade and investment facilitation and mutual recognition of professional qualifications. Using AICC information criterion, Hsiao et al. (2012) select six countries including Austria, Italy, Korea, Mexico, Norway and Singapore, and use OLS to construct the counterfactual GDP of Hong Kong if CEPA was not implemented.

Some preliminary results based on the OLS regression control method are provided in Section B of Supplementary Material. In particular, we replicate Hsiao et al. (2012)'s results with Hong Kong's actual GDP and its OLS prediction (the gap graph), and report it in Figure B.1. Moroever, Figure B.2 presents the point estimates of the treatment effects using RCM (based on OLS). While the estimated treatment effects remain positive throughout the post-treatment periods, we are unsure whether they are statistically significant, since no pointwise standard errors, confidence intervals or p-values are given in Hsiao et al. (2012).

We next revisit this dataset using the proposed QCM to study the effect of the economic integration between Hong Kong and mainland China. When implementing Random Forest (using R package `randomForest`) and QRF (using R package `quantregForest`), we again set the number of trees to be 1000, and turn off the bootstrap part of the algorithm. Since there are only 24 control units, we use the default value of 5 for the minimum size of terminal nodes. We also provide a Stata command qcm available froom SSC for easy implementation of QCM.

Figure 3.2 graphs Hong Kong's actual GDP and its Random Forest prediction (i.e., using Random Forest for point estimation, as we discussed in Remark 3.2), where the pre-treatment R^2 reaches 0.970. It is interesting to observe that while the Random Forest prediction has a better overall pre-treatment fit than RCM prediction (which is 0.931, see Figure B.1 in Supplementary Material), the former actually misses some of the deep trough and subsequent sharp rebounce following the political integration of Hong Kong with mainland China in 1997Q3. Since the event of political integration

with mainland China presumably only impacted Hong Kong to a large extent, the fluctuation of Hong Kong's GDP following 1997Q3 in Figure 3.2 is not supposed to be fully explained by other control countries or regions' GDP movements. In light of this, the near perfect RCM prediction following 1997Q3 appears to be overfit, which may reduce its ability to generalize to future unseen data.

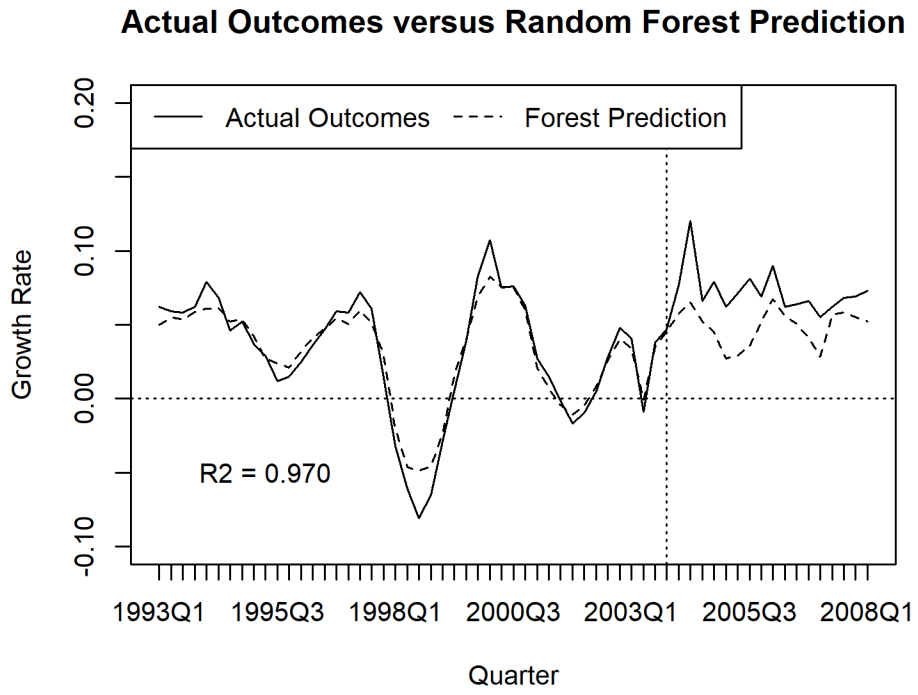


Figure 3.2: Actual Outcomes versus Random Forest Prediction

Figure 3.3 graphs the point estimates of treatment effects by Random Forest, as well as the 95% prediction intervals by QCM. It is clear from Figure 3.3 that only the treatment effect for the second period after the treatment (i.e., 2004Q2) is statistically significant at the 5% level, since its associated prediction interval does not contain zero; whereas prediction intervals for all other post-treatment periods contain zero. This is reminiscent of the effects of temporary boom in West Germany's GDP following the German reunification (Abadie et al., 2015). However, the effects of the economic integration with mainland China on Hong Kong's GDP remained in the positive territory afterwards, despite losing their economic and statistical significance over time.

In Figure 3.3, the point estimates of the mean treatment effects via Random Forest appear to be mostly centered in the QCM confidence intervals, which may not always be the case if the conditional

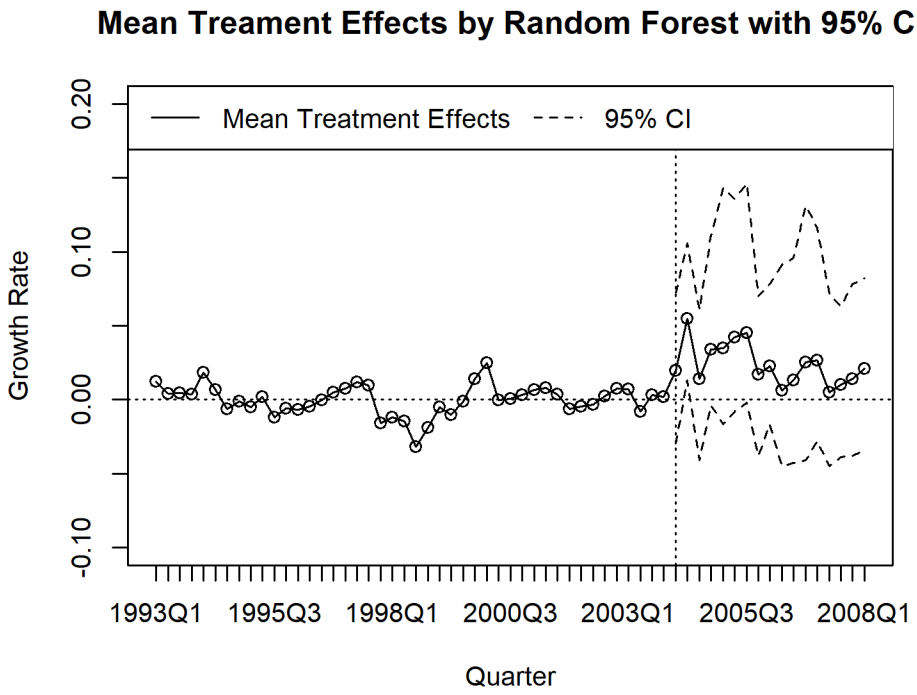


Figure 3.3: Mean Treatment Effects by Random Forest with 95% CI

distribution is not symmetric. Alternatively, one could use "median treatment effects" estimated by QRF at the 50% quantile as point estimates, as we discussed in Remark 3.3. The results are presented in Figure B.3 in Section B of Supplementary Material, which are very similar to Figure 3.3.

Detailed information behind Figure 3.3 and Figure B.3 are also presented in the Supplementary Material. As robustness checks, in Section B of Supplementary Material, we also conduct in-space and in-time placebo tests, which yield results consistent with QCM. Moreover, when we restrict the pre-treatment periods to avoid potential confounding events, the results from QCM are still robust, as reported in the Supplementary Material. Comparing to the empirical results based on the RCM, the proposed QRF method indicates smaller treatment effects of the CEPA on Hong Kong's economic growth.

3.7 Conclusion

In this paper, we study robust inference for treatment effects in panel data under SCM framework. We propose a simple way to construct pointwise confidence intervals for the treatment effects via QRF. As a nonparametric ensemble learning based on decision trees, the greatest strength of QRF lies in its robustness to heteroskedasticity, autocorrelation and model misspecification. Since this approach uses quantile regression and QRF in particular to construct a counterfactual control unit with its relevant quantiles, we call it QCM. Under some regularity conditions, we prove that the proposed method is asymptotically valid under our panel and time-series setting. Monte Carlo simulations show that QCM confidence intervals have excellent coverage probability close to the nominal rate even in small samples, which are robust to heteroskedasticity, autocorrelation, and model misspecification. We also revisit the case study of the economic integration between Hong Kong and mainland China to demonstrate the usefulness of QCM.

The basic idea of our proposed inference via QCM is straightforward for empirical practitioners. Moreover, QCM can be easily implemented by using forthcoming packages `qcm` in both R and Stata. We hope that practitioners would find QCM a reliable and robust approach of inference while estimating treatment effects for a single treated unit with panel data.

3.8 Technical Details

3.8.1 Approximating Rectangles

The proof of the consistency of QRF will be based on the technique of approximating rectangles developed by Wager and Walther (2016). For any data-dependent node $R(\mathbf{X}_0, \theta)$, Wager and Walther (2016) propose to use a set of rectangles to approximate it, where the rectangles are predetermined and do not depend on the data. This section introduces some basic results of approximating rectangles.

For a balanced tree, there are 2^k terminal nodes at splitting level k. For any positive integer k and $1 \leq l \leq 2^k$, the l-th terminal node R_l can be represented as $R_l = \bigotimes_{j=1}^{N} [r_{l,j}^-, r_{l,j}^+]$[15], where

[15]Note that in practice, $R_l = \bigotimes_{j=1}^{N} R_{l,j}$, where $R_{l,j}$ could be $[r_{l,j}^-, r_{l,j}^+]$, $(r_{l,j}^-, r_{l,j}^+]$, $[r_{l,j}^-, r_{l,j}^+)$ or $(r_{l,j}^-, r_{l,j}^+)$. For notational ease, we do not distinguish between all of these situations.

$0 \leq r_{l,j}^- < r_{l,j}^+ \leq 1$. For any node (rectangle) $R = \bigotimes_{j=1}^N [r_j^-, r_j^+]$, define the support of R as $\mathcal{S}(R_l)$, where

$$\mathcal{S}(R) = \left\{ j \in \{1, 2, \cdots, N\} : \text{either } r_j^- > 0 \text{ or } r_j^+ < 1 \right\}.$$

Intuitively, the support of a terminal node is the collection of all j's such that along the j-th covariate R was cut at least once.

Denote $s = \min\{N, k\}$. At splitting level k, there obviously holds $|\mathcal{S}(R_l)| \leq s$. Let $\mathcal{S} \subseteq \{1, 2, \cdots, N\}$ such that $|\mathcal{S}| = s$. According to Wager and Walther (2016), each terminal node R_l with support in $\mathcal{S}$ can be approximated by a set of rectangles $\mathcal{R}_{\mathcal{S},w,\varepsilon}$ in the sense that, for any terminal node R_l, if μ is the Lebesgue measure and $\mu(R_l) \geq w$, we can find $R^+, R^- \in \mathcal{R}_{\mathcal{S},w,\varepsilon}$ such that $R^- \subseteq R_l \subseteq R^+$, and

$$e^{-\varepsilon} \mu\left(R^+\right) \leq \mu(R_l) \leq e^{\varepsilon} \mu\left(R^-\right). \tag{3.19}$$

Consequently, let

$$\mathcal{R}_{N,k,w,\varepsilon} = \cup_{\mathcal{S} \subseteq \{1,2,\cdots,N\}, |\mathcal{S}|=s} \mathcal{R}_{\mathcal{S},w,\varepsilon}.$$

If $\min_{1 \leq l \leq 2^k} \mu(R_l) \geq w$, then any R_l can be approximated by $\mathcal{R}_{N,k,w,\varepsilon}$ in the sense of (3.19).

Wager and Walther (2016) demonstrate that if $|\mathcal{S}| = s$, then

$$|\mathcal{R}_{\mathcal{S},w,\varepsilon}| = \frac{1}{w} \left(\frac{8s^2}{\varepsilon^2} \left(1 + \log_2 \left[\frac{1}{w} \right] \right) \right)^s \cdot (1 + O(\varepsilon)).$$

When $\mu(R_l) \geq \xi^{-k}$ for all l, taking $w = \xi^{-k}$ and $\varepsilon = o(1)$ implies that for T_0 sufficiently large, there holds $|\mathcal{R}_{N,k,w,\varepsilon}| \leq 2 \begin{pmatrix} N \\ s \end{pmatrix} \cdot \xi^k \cdot \left(C s^2 k \varepsilon^{-2} \right)^s$. Tedious algebra leads to

$$\log |\mathcal{R}_{N,k,w,\varepsilon}| \leq Ck \left(\log N + \log k + \log \varepsilon^{-1} \right),$$

for any combination of N and k.

Finally, when $w = \xi^{-k}$ and $\varepsilon = o(1)$ hold, there holds

$$\min \{ \mu(R) : R \in \mathcal{R}_{N,k,w,\varepsilon} \} \geq (2\xi)^{-k}.$$

This is because, consider $R \in \mathcal{R}_{\mathcal{S},w,\varepsilon}$, according to Wager and Walther (2016), the rectangle has length at least $w 2^{\tau_j}$ along covariate j, where $j \in \mathcal{S}$, $\tau_j \in 0, 1, \cdots, \left[\log_2 w^{-1} \right]$ and $\sum_{j \in \mathcal{S}} \tau_j \geq$

$(s-1)\log_2\left(\frac{1}{w}\right)-s$. So the volume of R is at least[16]

$$\prod_{j\in\mathcal{S}}\left(w2^{\tau_j}\right)=w^s2^{\sum_{j\in\mathcal{S}}\tau_j}\geq w^s\left(\frac{1}{w}\right)^{s-1}2^{-s}=w2^{-\min\{N,k\}}$$
$$\geq\xi^{-k}2^{-k}=(2\xi)^{-k}.$$

3.8.2 Additional Lemmas

This section displays the auxiliary lemmas that will be useful in the proof of our main results, whose proof can be found in Section C of Supplementary Material to this paper.

Lemma 3.4. *Under Assumption 3.1–Assumption 3.4, taking $w=\xi^{-k}$ and $\varepsilon=\pi(k,N,T_0)$, then there exists a constant $C>0$ such that*

$$\lim_{T_0\to\infty}P\left[\sup_{R\in\mathcal{R}_{N,k,w,\varepsilon}}\sup_{y\in\mathbb{R}}\left|\frac{1}{\#R}\sum_{\mathbf{X}_t\in R}\mathbf{1}(Y_t\leq y)-F(y|R)\right|\leq C\cdot\pi(k,N,T_0)\right]=1.$$

Lemma 3.5. *Under Assumption 3.1–Assumption 3.4, if w and ε are taken as in Lemma 3.4, then there exists a constant $C>0$ such that*

$$\lim_{T_0\to\infty}P\left[\sup_{R_1\subseteq R_2\in\mathcal{R}_{N,k,w,\varepsilon}}\left|\frac{\#R_1-\#R_2}{\#R_2}-\frac{P(R_1)-P(R_2)}{P(R_2)}\right|\leq C\cdot\pi(k,N,T_0)\right]=1.$$

Lemma 3.6. *Under Assumption 3.1–Assumption 3.4, if w and ε are taken as in Lemma 3.4, then there exists a constant $C>0$ such that*

$$\lim_{T_0\to\infty}P\left[\sup_{\substack{R^-\subseteq R^+\in\mathcal{R}_{N,k,w,\varepsilon}\\ \mu(R^+)\leq e^{2\varepsilon}\mu(R^-)}}\left|\frac{\#R^+-\#R^-}{\#R^+}\right|\leq C\cdot\pi(k,N,T_0)\right]=1.$$

Lemma 3.7. *Suppose that Assumption 3.1–Assumption 3.4 hold and $|Y|\leq 1$ almost surely, then there holds*

$$\sup_{\mu(R)\geq\xi^{-(k-1)},i,\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\left|\widehat{\mathcal{I}(R,i,\lambda)}-\mathcal{I}(R,i,\lambda)\right|=O_p(\pi(k,N,T_0)).$$

[16]Note that in Wager and Walther (2016), the approximating rectangles have the form $R=\bigotimes_{j=1}^N[r_j^-,r_j^+]$, and moreover, taking j for an example, when $r_j^-+w2^{\tau_j}>1$, r_j^+ is truncated to 1. This implies that when r_j^- is close to 1, $\mu(R)<(2\xi)^{-k}$ may occur since $r_j^+-r_j^-<w2^{\tau_j}$. To deal with this problem, we can slightly enlarge R by expanding $[r_j^-,1]$ to $[r_j^-,r_j^-+w2^{\tau_j}]$. Then (3.19) still holds, and $\min\{\mu(R):R\in\mathcal{R}_{N,k,w,\varepsilon}\}\geq(2\xi)^{-k}$ holds at the same time.

Lemma 3.8. *Suppose that Assumption 3.1, Assumption 3.5 and Assumption 3.6 hold,* $|Y| \leq 1$ *holds almost surely, and* k *is fixed, then* $\inf_{a \leq \eta \leq 1} \delta(\eta) > 0$ *holds for any* $0 < a < 1$.

Lemma 3.9. *Suppose that Assumption 3.1 and Assumption 3.5 hold and* $|Y| \leq 1$ *holds almost surely, then for any rectangle* $R = \bigotimes_{i=1}^{N} [r_i^-, r_i^+]$ *and* $j \in \mathcal{Q}_1$*, we have*

$$\mathcal{I}(R, j, \lambda) \leq C \cdot \left(r_j^+ - r_j^-\right),$$

where C *is a constant independent of* R*,* j*, and* λ.

3.8.3 Proofs of Main Results

Proof of Theorem 3.1

Proof. Take w and ε as those in Lemma 3.4. According to Wager and Walther (2016), under Assumption 3.1-Assumption 3.4, for any terminal node R_l with $\mu(R_l) \geq w$, we can find approximating rectangles $R^+, R^- \in \mathcal{R}_{N,k,w,\varepsilon}$ such that $R^- \subseteq R_l \subseteq R^+$, and $e^{-\varepsilon}\mu(R^+) \leq \mu(R_l) \leq e^{\varepsilon}\mu(R^-)$. Denote such pair of R^+ and R^- as R_l^+ and R_l^-. Then we have

$$\begin{aligned}
&\sup_{\mathbf{X} \in \chi} \sup_{y \in \mathbb{R}} \left| \widehat{F}_\theta(y | \mathbf{X}) - F(y | \mathbf{X} \in R(\mathbf{X}, \theta)) \right| \\
&\leq \sup_l \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R_l} \sum_{\mathbf{X}_t \in R_l} \mathbf{1}(Y_t \leq y) - \frac{1}{\#R_l^+} \sum_{\mathbf{X}_t \in R_l^+} \mathbf{1}(Y_t \leq y) \right| (i) \\
&+ \sup_l \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R_l^+} \sum_{\mathbf{X}_t \in R_l^+} \mathbf{1}(Y_t \leq y) - F\left(y | \mathbf{X} \in R_l^+\right) \right| (ii) \\
&+ \sup_l \sup_{y \in \mathbb{R}} \left| F\left(y | \mathbf{X}_t \in R_l^+\right) - F(y | \mathbf{X} \in R_l) \right| (iii).
\end{aligned}$$

For (i), we have that

$$\begin{aligned}
(i) &\leq \sup_l \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R_l} \sum_{\mathbf{X}_t \in R_l} \mathbf{1}(Y_t \leq y) - \frac{1}{\#R_l^+} \sum_{\mathbf{X}_t \in R_l} \mathbf{1}(Y_t \leq y) \right| + \sup_l \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R_l^+} \sum_{\mathbf{X}_t \in \left(R_l^+ - R_l\right)} \mathbf{1}(Y_t \leq y) \right| \\
&\leq 2 \sup_l \left| \frac{\#R_l^+ - \#R_l}{\#R_l^+} \right| \leq 2 \sup_{\substack{R^- \subseteq R^+ \in \mathcal{R}_{N,k,w,\varepsilon} \\ \mu(R^+) \leq e^{2\varepsilon}\mu(R^-)}} \left| \frac{\#R^+ - \#R^-}{\#R^+} \right|.
\end{aligned}$$

For (ii), we have

$$(ii) \leq \sup_{R \in \mathcal{R}_{N,k,w,\varepsilon}} \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R} \sum_{\mathbf{X}_t \in R} \mathbf{1}\left(Y_t \leq y\right) - F\left(y \middle| \mathbf{X} \in R\right) \right|.$$

For (iii), we have for T_0 sufficiently large,

$$\begin{aligned}
(iii) &\leq \sup_l \sup_{y \in \mathbb{R}} \left| \frac{\int_{\mathbf{X} \in R_l} \int_{-\infty}^{y} f\left(Y, \mathbf{X}\right) dY d\mathbf{X}}{P\left(\mathbf{X} \in R_l^+\right)} - \frac{\int_{\mathbf{X} \in R_l} \int_{-\infty}^{y} f\left(Y, \mathbf{X}\right) dY d\mathbf{X}}{P\left(\mathbf{X} \in R_l\right)} \right| \\
&+ \sup_l \sup_{y \in \mathbb{R}} \left| \frac{\int_{\mathbf{X} \in \left(R_l^+ - R_l\right)} \int_{-\infty}^{y} f\left(Y, \mathbf{X}\right) dY d\mathbf{X}}{P\left(\mathbf{X} \in R_l^+\right)} \right| \leq \sup_l \sup_{y \in \mathbb{R}} 2 \left| \frac{P\left(\mathbf{X} \in R_l^+\right) - P\left(\mathbf{X} \in R_l\right)}{P\left(\mathbf{X} \in R_l^+\right)} \right| \\
&\leq \sup_l \sup_{y \in \mathbb{R}} 2 \left| \frac{P\left(\mathbf{X} \in R_l^+\right) - P\left(\mathbf{X} \in R_l^-\right)}{P\left(\mathbf{X} \in R_l^+\right)} \right| \leq 2\zeta^2 \left(1 - e^{-2\varepsilon}\right) \leq 8\zeta^2 \varepsilon.
\end{aligned}$$

The above implies that for T_0 sufficiently large,

$$\begin{aligned}
&\sup_{\mathbf{X} \in \chi} \sup_{y \in \mathbb{R}} \left| \mathbb{E}_\theta \widehat{F}_\theta\left(y \middle| \mathbf{X}\right) - \mathbb{E}_\theta F\left(y \middle| \mathbf{X}_t \in R\left(\mathbf{X}, \theta\right)\right) \right| \\
&\leq 2 \sup_{\substack{R^- \subseteq R^+ \in \mathcal{R}_{N,k,w,\varepsilon} \\ \mu(R^+) \leq e^{2\varepsilon} \mu(R^-)}} \left| \frac{\#R^+ - \#R^-}{\#R^+} \right| + \sup_{R \in \mathcal{R}_{N,k,w,\varepsilon}} \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R} \sum_{\mathbf{X}_t \in R} \mathbf{1}\left(Y_t \leq y\right) - F\left(y \middle| \mathbf{X}_t \in R\right) \right| + 8\zeta^2 \varepsilon.
\end{aligned}$$

From Lemma 3.4, Lemma 3.5, and Lemma 3.6, let $C = 2C_3 + C_2 + 8\zeta^2$, we have

$$\begin{aligned}
&P\left[\sup_{\mathbf{x} \in \chi} \sup_{y \in \mathbb{R}} \left| \mathbb{E}_\theta \widehat{F}_\theta\left(y \middle| \mathbf{X}\right) - \mathbb{E}_\theta F\left(y \middle| \mathbf{X}_t \in R\left(\mathbf{X}, \theta\right)\right) \right| > C \cdot \pi\left(k, N, T_0\right) \right] \\
&\leq P\left[2 \sup_{\substack{R^- \subseteq R^+ \in \mathcal{R}_{N,k,w,\varepsilon} \\ \mu(R^+) \leq e^{2\varepsilon} \mu(R^-)}} \left| \frac{\#R^+ - \#R^-}{\#R^+} \right| > 2C_3 \pi\left(k, N, T_0\right) \right] \\
&+ P\left[\sup_{R \in \mathcal{R}_{N,k,w,\varepsilon}} \sup_{y \in \mathbb{R}} \left| \frac{1}{\#R} \sum_{\mathbf{X}_t \in R} I\left(Y_t \leq y\right) - F\left(y \middle| \mathbf{X}_t \in R\right) \right| > C_2 \pi\left(k, N, T_0\right) \right] \to 0.
\end{aligned}$$

This finishes the proof of Theorem 3.1. □

Proof of Theorem 3.2

Proof. Note that under Assumption 3.1, we have $\zeta^{-1} \leq f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right) \leq \zeta$ holds for any $\mathbf{X} \in \chi$. For $F\left(y \middle| R\left(\mathbf{X}_0, \theta\right)\right)$, we have that

$$\begin{aligned}
F\left(y \middle| R\left(\mathbf{X}_0, \theta\right)\right) &= \frac{\int_{\mathbf{X} \in R(\mathbf{X}_0, \theta)} \int_{-\infty}^{y} f_{Y|\mathbf{X}}\left(Y|\mathbf{X}\right) f_{\mathbf{X}}\left(\mathbf{X}\right) dY d\mathbf{X}}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)} \\
&= \frac{\int_{\mathbf{X} \in R(\mathbf{X}_0, \theta)} f_{\mathbf{X}}\left(\mathbf{X}\right) d\mathbf{X} \int_{-\infty}^{y} f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}]_{\mathcal{Q}}\right) dY}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)} \\
&= \frac{\int_{\mathbf{X} \in R(\mathbf{X}_0, \theta)} f_{\mathbf{X}}\left(\mathbf{X}\right) d\mathbf{X} \int_{-\infty}^{y} f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right) dY}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)} \\
&+ \frac{\int_{\mathbf{X} \in R(\mathbf{X}_0, \theta)} f_{\mathbf{X}}\left(\mathbf{X}\right) d\mathbf{X} \int_{-\infty}^{y} \left\{f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}]_{\mathcal{Q}}\right) - f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right)\right\} dY}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)}
\end{aligned}$$

For the first term on the RHS of the last equality, we have

$$\begin{aligned}
\frac{\int_{\mathbf{X} \in R(\mathbf{X}_0, \theta)} f_{\mathbf{X}}\left(\mathbf{X}\right) d\mathbf{X} \int_{-\infty}^{y} f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right) dY}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)} &= \frac{F\left(y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right) \cdot P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)}{P\left(\mathbf{X} \in R\left(\mathbf{X}_0, \theta\right)\right)} \\
&= F\left(y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right).
\end{aligned}$$

The remaining task is to obtain an upper bound for the second term. Note that

$$\begin{aligned}
\left|f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}]_{\mathcal{Q}}\right) - f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y \middle| [\mathbf{X}_0]_{\mathcal{Q}}\right)\right| &= \left|\frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} - \frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)}\right| \\
&\leq \left|\frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} - \frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)}\right| + \left|\frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} - \frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)}\right|.
\end{aligned}$$

For the first term on the RHS of the inequality, we have

$$\left|\frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} - \frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)}\right| \leq \frac{L\left(Y\right) \cdot \left\|[\mathbf{X}]_{\mathcal{Q}} - [\mathbf{X}_0]_{\mathcal{Q}}\right\|}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} \leq \zeta L\left(Y\right) \cdot \left\|[\mathbf{X}]_{\mathcal{Q}} - [\mathbf{X}_0]_{\mathcal{Q}}\right\|.$$

For the second term, we have

$$\begin{aligned}
\left|\frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} - \frac{f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)}\right| &\leq f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right) \cdot \frac{\left|f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right) - f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)\right|}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right) f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}]_{\mathcal{Q}}\right)} \\
&\leq \frac{L\zeta f_{\mathcal{Q}}\left(Y, [\mathbf{X}_0]_{\mathcal{Q}}\right) \cdot \left\|[\mathbf{X}]_{\mathcal{Q}} - [\mathbf{X}_0]_{\mathcal{Q}}\right\|}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)}.
\end{aligned}$$

Since $\left\|[\mathbf{X}]_{\mathcal{Q}}-[\mathbf{X}_0]_{\mathcal{Q}}\right\| \leq \operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)$, we have

$$
\begin{aligned}
&\left|\frac{\int_{\mathbf{X}\in R(\mathbf{X}_0,\theta)} f_{[\mathbf{X}]_{\mathcal{Q}}}(\mathbf{X})\,d\mathbf{X}\int_{-\infty}^{y}\left\{f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y\left|[\mathbf{X}]_{\mathcal{Q}}\right.\right)-f_{Y|[\mathbf{X}]_{\mathcal{Q}}}\left(Y\left|[\mathbf{X}_0]_{\mathcal{Q}}\right.\right)\right\}dY}{P(\mathbf{X}\in R(\mathbf{X}_0,\theta))}\right| \\
&\leq \frac{\int_{\mathbf{X}\in R(\mathbf{X}_0,\theta)} f_{[\mathbf{X}]_{\mathcal{Q}}}(\mathbf{X})\,d\mathbf{X}\int_{-\infty}^{y}\zeta L(Y)\cdot\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)dY}{P(\mathbf{X}\in R(\mathbf{X}_0,\theta))} \\
&+\frac{\int_{\mathbf{X}\in R(\mathbf{X}_0,\theta)} f_{[\mathbf{X}]_{\mathcal{Q}}}(\mathbf{X})\,d\mathbf{X}\int_{-\infty}^{y}\frac{Lf_{\mathcal{Q}}\left(Y,[\mathbf{X}_0]_{\mathcal{Q}}\right)\cdot\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)}dY}{P(\mathbf{X}\in R(\mathbf{X}_0,\theta))} \\
&\leq \frac{\int_{\mathbf{X}\in R(\mathbf{X}_0,\theta)} f_{[\mathbf{X}]_{\mathcal{Q}}}(\mathbf{X})\,d\mathbf{X}\int_{-\infty}^{\infty}\zeta L(Y)\cdot\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)dY}{P(\mathbf{X}\in R(\mathbf{X}_0,\theta))} \\
&+\frac{\int_{\mathbf{X}\in R(\mathbf{X}_0,\theta)} f_{[\mathbf{X}]_{\mathcal{Q}}}(\mathbf{X})\,d\mathbf{X}\int_{-\infty}^{\infty}\frac{Lf_{\mathcal{Q}}\left(Y,[\mathbf{X}_0]_{\mathcal{Q}}\right)\cdot\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)}{f_{[\mathbf{X}]_{\mathcal{Q}}}\left([\mathbf{X}_0]_{\mathcal{Q}}\right)}dY}{P(\mathbf{X}\in R(\mathbf{X}_0,\theta))} \\
&\leq \zeta L\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)+\zeta L\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right)=C\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right).
\end{aligned}
$$

As a result, we have

$$
\left|F\left(y\left|R(\mathbf{X}_0,\theta)\right.\right)-F\left(y\left|\mathbf{X}_0\right.\right)\right| \leq C\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right).
$$

So

$$
\begin{aligned}
\left|\mathbb{E}_\theta F\left(y\left|R(\mathbf{X}_0,\theta)\right.\right)-F\left(y\left|\mathbf{X}_0\right.\right)\right| &\leq \mathbb{E}_\theta\left|F\left(y\left|\mathbf{X}\in R(\mathbf{X}_0,\theta)\right.\right)-F\left(y\left|\mathbf{X}_0\right.\right)\right| \\
&\leq C\mathbb{E}_\theta\operatorname{diam}\left([R(\mathbf{X}_0,\theta)]_{\mathcal{Q}}\right).
\end{aligned}
$$

Taking supreme for both sides with respect to $\mathbf{X}_0$, this finishes the proof of Theorem 3.2. □

Proof of Lemma 3.1

Proof. We only need to show that

$$
\lim_{T_0\to\infty} P\left[\inf_{\mathbf{X}_0\in\mathcal{X}:\Psi(\mathbf{X}_0,k,d)\neq\emptyset}\ \inf_{\theta\in\Psi(\mathbf{X}_0,k,d)}\sum_{j\in\mathcal{Q}_1}\mathcal{N}(\mathbf{X}_0,\theta,k,j)\leq d-1\right]=0.
$$

Note that

$$\left\{\inf_{\mathbf{X}_0\in\mathcal{X}:\Psi(\mathbf{X}_0,k,d)\neq\emptyset}\inf_{\theta\in\Psi(k,d)}\sum_{j\in\mathcal{Q}_1}\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)\leq d-1\right\}$$
$$\subseteq\left\{\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\left[\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(\widehat{R,i,\lambda}\right)-\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(\widehat{R,j,\lambda}\right)\right]\geq0\right\}.$$

Since

$$\begin{aligned}&\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\left[\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(\widehat{R,i,\lambda}\right)-\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(\widehat{R,j,\lambda}\right)\right]\\&\leq\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\left[\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,i,\lambda\right)-\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,j,\lambda\right)\right]\\&+\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\left|\mathcal{I}\left(\widehat{R,i,\lambda}\right)-\mathcal{I}\left(R,i,\lambda\right)\right|\\&+\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\left|\mathcal{I}\left(\widehat{R,j,\lambda}\right)-\mathcal{I}\left(R,j,\lambda\right)\right|.\end{aligned}$$

According to Lemma 3.8 and Assumption 3.7, we have

$$\begin{aligned}&\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\left[\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,i,\lambda\right)-\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,j,\lambda\right)\right]\\&\leq-\left(1-\omega\right)\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,j,\lambda\right)\leq-\left(1-\omega\right)\inf_{\eta\geq\xi^{-(k-1)}}\delta\left(\eta\right).\end{aligned}$$

According to Lemma 3.7, we have

$$\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\left|\mathcal{I}\left(\widehat{R,i,\lambda}\right)-\mathcal{I}\left(R,i,\lambda\right)\right|=O_p\left(\pi\left(k,N,T_0\right)\right),$$

and

$$\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\left|\mathcal{I}\left(\widehat{R,j,\lambda}\right)-\mathcal{I}\left(R,j,\lambda\right)\right|=O_p\left(\pi\left(k,N,T_0\right)\right).$$

The above implies that

$$\begin{aligned}\lim_{T_0\to\infty}P\Bigg[&\sup_{R\subseteq\chi:\mu(R)\geq\xi^{-(k-1)}}\left[\sup_{i\notin\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,i,\lambda\right)-\sup_{j\in\mathcal{Q}_1}\sup_{\xi^{-1}\leq\lambda\leq1-\xi^{-1}}\mathcal{I}\left(R,j,\lambda\right)\right]\\&>-\frac{1}{2}\left(1-\omega\right)\inf_{\eta\geq\xi^{-(k-1)}}\delta\left(\eta\right)\Bigg]=0.\end{aligned}$$

This leads to the desired result. □

Proof of Lemma 3.2

Proof. First note that since the total number of splits is fixed at k, with probability going to 1, we only split on the covariates within set $\mathcal{Q}_1$ when all the covariates in $\mathcal{Q}_1$ are simultaneously selected as candidates. For any fixed $\mathbf{X}_0$ such that $\Psi(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\}) \neq \emptyset$ and any fixed $\theta \in \Psi(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\})$, since the total number of rounds in which all covariates in $\mathcal{Q}_1$ are simultaneously selected as candidates is $|\mathcal{Q}_1| \cdot \max\{d, d^*\}$, then at least one covariate $j_1 \in \mathcal{Q}_1$ is chosen and split for no less than d^* times within the rounds in which all the covariates in $\mathcal{Q}_1$ are selected as candidates. Suppose there is some covariate $j_2 \in \mathcal{Q}$ such that it was split with less than d times, then $\sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j_2, \lambda)$ is lower bounded by $\inf_{\eta \geq \xi^{-(d-1)}} \delta(\eta)$ according to Lemma 3.8. Consider the round where all the covariates in $\mathcal{Q}_1$ are selected as candidates and the last split of j_1 takes place. In this round, $\sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j_1, \lambda)$ is upper bounded by $C \cdot \left(1 - \xi^{-1}\right)^{d^*-1}$ according to Lemma 3.9, which is strictly smaller than $\frac{1}{2} \inf_{\eta \geq \xi^{-(d-1)}} \delta(\eta)$. So the difference between $\sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j_1, \lambda)$ and $\sup_{\xi^{-1} \leq \lambda \leq 1-\xi^{-1}} \mathcal{I}(R, j_2, \lambda)$ is at least $\frac{1}{2} \inf_{\eta \geq \xi^{-(d-1)}} \delta(\eta)$. Moreover, Lemma 3.7 implies the estimation error degenerates to zero uniformly with respect to all the nodes that may appear in the first k rounds of splits. So j_2 should be split instead of j_1 with probability going to 1, which leads to a contradiction. Note that the above argument does not depend on the specific $\mathbf{X}_0$ or θ, so we prove the result. □

Proof of Lemma 3.3

Proof. We first prove the result under diverging N. Note that for any fixed k, Lemma 3.2 leads to

$$\lim_{T_0 \to \infty} P\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\}\right) \subseteq \left\{\theta : \min_{j \in \mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0, \theta, k, j\right) \geq d\right\} \text{ for all } \mathbf{X}_0\right] = 1,$$

which implies that

$$\lim_{T_0 \to \infty} P\left[P_\theta\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\}\right)\right] \leq P_\theta\left[\min_{j \in \mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0, \theta, k, j\right) \geq d\right] \text{ for all } \mathbf{X}_0\right] = 1,$$

and

$$\lim_{T_0 \to \infty} P\left[\inf_{\mathbf{X}_0 \in \mathcal{X}} P_\theta\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1| \cdot \max\{d, d^*\}\right)\right] \leq \inf_{\mathbf{X}_0 \in \mathcal{X}} P_\theta\left[\min_{j \in \mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0, \theta, k, j\right) \geq d\right]\right] = 1.$$

Note that in each round of random feature selection, a total of m_{try} features are selected. So in a single round, the probability of drawing all covariates in $\mathcal{Q}_1$ simultaneously is given by

$$\begin{aligned}\frac{C_{N-|\mathcal{Q}_1|}^{m_{try}-|\mathcal{Q}_1|}}{C_N^{m_{try}}} &= \left(1-\frac{|\mathcal{Q}_1|}{N}\right)\left(1-\frac{|\mathcal{Q}_1|}{N-1}\right)\cdots\left(1-\frac{|\mathcal{Q}_1|}{m_{try}}\right) \\ &\geq \left(1-\frac{|\mathcal{Q}_1|}{m_{try}}\right)^{N-m_{try}} = \left(\left(1-\frac{|\mathcal{Q}_1|}{m_{try}}\right)^{m_{try}}\right)^{\frac{N-m_{try}}{m_{try}}} \geq e^{\frac{1}{2}\left(\liminf_{N\to\infty}\left(1-\frac{N}{m_{try}}\right)\right)} \equiv \underline{P} > 0\end{aligned}$$

for N large. Then for N sufficiently large, we have that

$$P_\theta\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1|\cdot\max\{d,d^*\}\right)\right] \geq 1 - \sum_{j=0}^{|\mathcal{Q}_1|\cdot\max\{d,d^*\}-1} C_k^j(1-\underline{P})^{k-j}\underline{P}^j.$$

Note that the RHS does not depend on $\mathbf{X}_0$, so we have that

$$\inf_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1|\cdot\max\{d,d^*\}\right)\right] \geq 1 - \sum_{j=0}^{|\mathcal{Q}_1|\cdot\max\{d,d^*\}-1} C_k^j(1-\underline{P})^{k-j}\underline{P}^j.$$

Take limit with respect to k for both sides, we have that

$$\begin{aligned}\lim_{k\to\infty}\inf_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left[\Psi\left(\mathbf{X}_0, k, |\mathcal{Q}_1|\cdot\max\{d,d^*\}\right)\right] &\geq 1 - \lim_{k\to\infty}\sum_{j=0}^{|\mathcal{Q}_1|\cdot\max\{d,d^*\}-1} C_k^j(1-\underline{P})^{k-j}\underline{P}^j \\ &= 1.\end{aligned}$$

Then for any $0 < c < 1$, for k that is sufficiently large, we have that

$$\lim_{T_0\to\infty} P\left[\inf_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left[\min_{j\in\mathcal{Q}}\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)\geq d\right] > c\right] = 1$$

holds. As a result, we have

$$p\lim_{T_0\to\infty}\inf_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left[\min_{j\in\mathcal{Q}}\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)\geq d\right] = 1.$$

Now we prove the result under fixed N, in which case $\mathcal{Q}_1 = \{1, 2, \cdots, N\}$. We will only consider $m_{try} > 1$, otherwise the proof is trivial. We first assume that we observe $\mathcal{I}(R, j, \lambda)$ directly. Note that showing the result is equivalent to showing that

$$\lim_{k\to\infty}\sup_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left[\min_{j\in\mathcal{Q}}\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) < d\right] = 0$$

for any fixed d. Note that

$$P_\theta\left[\min_{j\in\mathcal{Q}}\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)<d\right]\leq\sum_{j\in\mathcal{Q}}P_\theta\left[\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)<d\right],$$

so we only need to show that

$$\lim_{k\to\infty}\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right)<d\right]=0$$

for each $j\in\mathcal{Q}$. Suppose the above result does not hold, then for some fixed d and some $j^*\in\mathcal{Q}$, there holds

$$\liminf_{k\to\infty}\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\right]>0.$$

Define $\Phi(\mathbf{X}_0,k,j^*,N\cdot s)$ as the collection of θ such that any covariate $j\neq j^*$ is simultaneously selected with j^* with no less than $N\cdot s$ times. Then for any $\theta\in\Phi(\mathbf{X}_0,k,j^*,N\cdot s)$, at least one covariate is split with more than s times when it is jointly selected with covariate j^* (such covariate can be j^* itself). Now let $s=\max\{d,d^*\}$, where d^* is specified in the proof of Lemma 3.2. If for such θ, covariate j^* is split with no less than $\max\{d,d^*\}$ times, then such θ can not be contained in set $\{\theta:\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\}$. If for such θ, some covariate other than j^* is split with no less than $\max\{d,d^*\}$ times, then consider the last round where covariate j is selected with j^* and is split over. If at that point covariate j is split with less than d times, then $\sup_\lambda\mathcal{I}(R,j,\lambda)\leq 1/2\inf_{\eta\geq\xi^{-(d-1)}}\delta(\eta)$ while $\sup_\lambda\mathcal{I}(R,j^*,\lambda)\geq\inf_{\eta\geq\xi^{-(d-1)}}\delta(\eta)$, implying that in that round j^*, instead of j, should be split, which leads to a contradiction. This also implies that θ can not be in the set $\{\theta:\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\}$. So together $\{\theta:\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\}\bigcap\Phi(\mathbf{X}_0,k,j^*,N\cdot\max\{d,d^*\})=\emptyset$. Then

$$\begin{aligned}1\geq\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\{\text{The Collection of All }\theta\}\right]\geq&\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\Phi(\mathbf{X}_0,k,j^*,N\cdot\max\{d,d^*\})\right]\\&+\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\right].\end{aligned}$$

Taking lim inf for both sides with respect to k, we have that

$$\liminf_{k\to\infty}\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\Phi(\mathbf{X}_0,k,j^*,N\cdot\max\{d,d^*\})\right]=1$$

and this leads to a contradiction if $\liminf_{k\to\infty}\sup_{\mathbf{X}_0\in\mathcal{X}}P_\theta\left[\mathcal{N}\left(\mathbf{X}_0,\theta,k,j^*\right)<d\right]>0$. So it can not happen. Note that above proof requires that we observe $\mathcal{I}(R,j,\lambda)$, but again, Lemma 3.7 implies

that the estimation error is uniform bounded for any fixed k, this finishes the proof. □

Proof of Theorem 3.3

Proof. According to Theorem 3.1 and Theorem 3.2, it remains to show that

$$\sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \to_p 0.$$

Since

$$\begin{aligned}
&\sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \\
&\leq \sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \left\{ \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \middle| \min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) \geq d \right\} P_\theta\left(\min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) \geq d\right) \\
&+ \sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \left\{ \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \middle| \min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) < d \right\} P_\theta\left(\min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) < d\right) \\
&\leq \left(1-\xi^{-1}\right)^d + \sup_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left(\min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) < d\right),
\end{aligned}$$

and according to Lemma 3.3, we have that $\sup_{\mathbf{X}_0\in\mathcal{X}} P_\theta\left(\min_{j\in\mathcal{Q}} \mathcal{N}\left(\mathbf{X}_0,\theta,k,j\right) < d\right) \to_p 0$ as $T_0\to\infty$ and $k\to\infty$, so

$$\lim_{T_0,k\to\infty} P\left[\sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \leq 2\left(1-\xi^{-1}\right)^d\right] = 1.$$

Since this holds for any d, we have $\sup_{\mathbf{X}_0\in\mathcal{X}} \mathbb{E}_\theta \mathrm{diam}\left(\left[R\left(\mathbf{X}_0,\theta\right)\right]_{\mathcal{Q}}\right) \to_p 0$, and hence the uniform consistency result is proved.

To prove that the conditional quantile estimator is uniformly consistent, we first show that

$$\sup_{\mathbf{X}_0\in\mathcal{X},0\leq\alpha\leq 1} \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - \alpha\right| \to_p 0.$$

Suppose that the above does not hold, we can find a sequence of $\mathbf{X}_0^{T_0}$ and α^{T_0} and some positive constant υ such that

$$\liminf_{n\to\infty} P\left(\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0})|\mathbf{X}_0^{T_0}) - \alpha^{T_0} \geq \upsilon\right) > 0.$$

Note that according to the definition of quantile function, $\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0}) - \varrho|\mathbf{X}_0^{T_0}) < \alpha^{T_0}$ for

arbitrary small $\varrho > 0$. Then

$$\begin{aligned}
\upsilon &< \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0}) - \varrho|\mathbf{X}_0^{T_0}) - \mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0})|\mathbf{X}_0^{T_0})\right| \\
&\leq \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0}) - \varrho|\mathbf{X}_0^{T_0}) - F(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0}) - \varrho|\mathbf{X}_0^{T_0})\right| \\
&+ \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0})|\mathbf{X}_0^{T_0}) - F(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0})|\mathbf{X}_0^{T_0})\right| \\
&+ \left|F(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0}) - \varrho|\mathbf{X}_0^{T_0}) - F(\widehat{Q}_Y(\alpha^{T_0}|\mathbf{X}_0^{T_0})|\mathbf{X}_0^{T_0})\right| \\
&\leq 2 \sup_{\mathbf{X}_0 \in \mathcal{X}, y} \left|\mathbb{E}_\theta \widehat{F}_\theta\left(y|\,\mathbf{X}_0\right) - F\left(y|\,\mathbf{X}_0\right)\right| + \sup_{\mathbf{X}_0 \in \mathcal{X}, y} F_y\left(y|\,\mathbf{X}_0\right) \cdot \varrho
\end{aligned}$$

Obviously, we can choose ϱ small such that $\sup_{\mathbf{X}_0 \in \mathcal{X}, y} F_y\left(y|\,\mathbf{X}_0\right) \cdot \varrho < \upsilon/2$. Then the above implies that $\sup_{\mathbf{X}_0 \in \mathcal{X}, y} \left|\mathbb{E}_\theta \widehat{F}_\theta\left(y|\,\mathbf{X}_0\right) - F\left(y|\,\mathbf{X}_0\right)\right| > \upsilon/4$. But such event has probability going to zero according to the uniform consistency of the conditional CDF estimator. This leads to a contradiction and hence proves our result.

Now we prove our theorem. For any a and positive constant c, define $(a)_c = a$ if $-c \leq a \leq c$, $(a)_c = -c$ if $a < -c$ and $(a)_c = c$ if $a > c$. We have that,

$$\begin{aligned}
&\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - \alpha = \mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F(Q_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) \\
&= \mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) + F(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F((\widehat{Q}_Y(\alpha|\mathbf{X}_0))_c|\mathbf{X}_0) \\
&+ F((\widehat{Q}_Y(\alpha|\mathbf{X}_0))_c|\mathbf{X}_0) - F((Q_Y(\alpha|\mathbf{X}_0)_c|\mathbf{X}_0) + F((Q_Y(\alpha|\mathbf{X}_0))_c|\mathbf{X}_0) - F(Q_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0).
\end{aligned}$$

Define $\underline{L}_c = \inf_{\mathbf{X}_0 \in \mathcal{X}, -c \leq y \leq c} F_y(y|\mathbf{X}_0)$, which is strictly positive for any positive $c < 1$, then

$$\begin{aligned}
&\sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \leq \alpha \leq 1} \left|(\widehat{Q}_Y(\alpha|\mathbf{X}_0))_c - (Q_Y(\alpha|\mathbf{X}_0))_c\right| \\
&\leq \underline{L}_c^{-1} \left\{ \sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \leq \alpha \leq 1} \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0)\right| \right. \\
&\left. + 2 \left(\sup_{\mathbf{X}_0 \in \mathcal{X}, |y| \geq c} F_y(y|\mathbf{X}_0) \right) \cdot (1 - c) + \left(\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - \alpha \right) \right\},
\end{aligned}$$

where the RHS comes from the fact that $|(a)_c - a| \leq (1 - c)$ for any $|a| \leq 1$ and $0 < c < 1$. Since $\left(\inf_{\mathbf{X}_0 \in \mathcal{X}, -c \leq y \leq c} F_y\left(y|\,\mathbf{X}_0\right)\right)^{-1} \cdot (1 - c) \cdot \sup_{\mathbf{X}_0 \in \mathcal{X}, |y| \geq c} F_y\left(y|\,\mathbf{X}_0\right) \to 0$ for $c \to 1$, we have that for any ε,

$$2\underline{L}_c^{-1} \left(\sup_{\mathbf{X}_0 \in \mathcal{X}, |y| \geq c} F_y(y|\mathbf{X}_0) \right) \cdot (1 - c) < \varepsilon/2$$

for c sufficiently close to 1. Since $\sup_{\mathbf{X}_0 \in \mathcal{X}, \alpha} \left|\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0)\right| \to_p 0$ and

$\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - \alpha \to_p 0$, we have that for the above mentioned fixed c, with probability going to 1,

$$\underline{L}_c^{-1} \cdot \sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \le \alpha \le 1} \left| \mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - F(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) \right| < \varepsilon/4,$$

and

$$\underline{L}_c^{-1} \cdot \left(\mathbb{E}_\theta \widehat{F}_\theta(\widehat{Q}_Y(\alpha|\mathbf{X}_0)|\mathbf{X}_0) - \alpha \right) < \varepsilon/4.$$

So $\sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \le \alpha \le 1} \left| (\widehat{Q}_Y(\alpha|\mathbf{X}_0))_c - (Q_Y(\alpha|\mathbf{X}_0)_c \right|$ is bounded by ε for c sufficiently close to 1 with probability going to 1. Finally, note that

$$\sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \le \alpha \le 1} \left| \widehat{Q}_Y(\alpha|\mathbf{X}_0) - Q_Y(\alpha|\mathbf{X}_0 \right| \le \left| (\widehat{Q}_Y(\alpha|\mathbf{X}_0))_c - (Q_Y(\alpha|\mathbf{X}_0)_c \right| + 2(1-c).$$

So for c sufficiently close to 1 with probability going to 1, there holds

$$\sup_{\mathbf{X}_0 \in \mathcal{X}, 0 \le \alpha \le 1} \left| \widehat{Q}_Y(\alpha|\mathbf{X}_0) - Q_Y(\alpha|\mathbf{X}_0) \right| \le \varepsilon + 2(1-c).$$

Finally, since ε and c are both arbitrary, we prove the result. □

3.8.4 Proof of Proposition 3.1

We first list a lemma, whose proof can be found in the proof of Lemma 3.7 in the Supplementary Material.

Lemma 3.10. *Suppose that Assumption 3.1, Assumption 3.2, and Assumption 3.4 hold, we have that*

$$\sup_{R' \subseteq R, \mu(R) \ge \xi^{-(k-1)}, \mu(R') \ge \xi^{-k}} \left| \frac{\#R'}{\#R} - \frac{P(R')}{P(R)} \right| = O_p\left(\pi(k, N, T_0)\right).$$

Based on Lemma 3.10, we can prove the Proposition 3.1 in the main context.

Proof of Proposition 3.1. Let's consider any set $R = \bigotimes_{i=1}^N [r_i^-, r_i^+]$ such that $\mu(R) \ge \xi^{-(k-1)}$. Define $\underline{R}_j = R \cap \{X_j : r_i^- \le X_j \le \left(1 - \xi^{-1}\right) r_j^- + \xi^{-1} r_j^+\}$. Obviously, there hold $\mu(\underline{R}_j) \ge \xi^{-k}$ and $P(\underline{R}_j)/P(R) \le \zeta^2 \xi^{-1}$. According to Lemma 3.10, with probability going to 1, there holds

$$\frac{\#\underline{R}_j}{\#R} \le \frac{P(\underline{R}_j)}{P(R)} + 0.1\zeta^2\xi^{-1} \le 1.1\zeta^2\xi^{-1}.$$

Then since $1.1\zeta^2\xi^{-1} < \widetilde{\xi}^{-1}$, we have that $\#\underline{R}_j/\#R < \widetilde{\xi}^{-1}$, which indicates that to ensure that

the child node contains at least $\widetilde{\xi}^{-1}$ proportion of the observations in the parent node R, the splitting point along direction j can not be smaller than $\left(1-\xi^{-1}\right) r_j^- + \xi^{-1} r_j^+$. Similar arguments directly lead to that the splitting point along direction j can not be larger than $\xi^{-1} r_j^- + \left(1-\xi^{-1}\right) r_j^+$, either. It then remains to use induction to show that starting from χ, for any parent node $R = \bigotimes_{i=1}^N [r_i^-, r_i^+]$ and splitting direction j, the splitting point lies within the interval $\left[\left(1-\xi^{-1}\right) r_j^- + \xi^{-1} r_j^+,\ \xi^{-1} r_j^- + \left(1-\xi^{-1}\right) r_j^+\right]$. Then (A) is proved. The proof of (B) can be done similarly, so is omitted.

□

Bibliography

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Alekh Agarwal, Sham Kakade, Nikos Karampatziakis, Le Song, and Gregory Valiant. Least squares revisited: Scalable approaches for multi-class prediction. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2014.

Hyungtaik Ahn, Hidehiko Ichimura, James L Powell, and Paul A Ruud. Simple estimators for invertible index models. *Journal of Business & Economic Statistics*, 36(1):1–10, 2018.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

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

Alexandre Belloni and Victor Chernozhukov. â1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82 – 130, 2011.

Alexandre Belloni, Victor Chernozhukov, and Kengo Kato. High-dimensional quantile regression. In *Handbook of quantile regression*, pages 253–272. Chapman and Hall/CRC, 2017.

Eli Ben-Michael, Avi Feller, and Jesse Rothstein. The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803, 2021.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1):1063–1095, 2012.

Herman J Bierens. Consistency and asymptotic normality of sieve ml estimators under low-level conditions. *Econometric Theory*, 30(5):1021–1076, 2014.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman. Consistency for a simple model of random forests. 2004.

Carlos Carvalho, Ricardo Masini, and Marcelo C Medeiros. Arco: an artificial counterfactual approach for high-dimensional panel time-series data. *Journal of econometrics*, 207(2):352–380, 2018.

Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021.

M.D. Cattaneo, M. Jansson, and W.K. Newey. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34:277–301, 2018a.

M.D. Cattaneo, M. Jansson, and W.K. Newey. Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association*, 113(523):1350–1361, 2018b.

Christopher Cavanagh and Robert P Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–381, 1998.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

Victor Chernozhukov, Kaspar Wuthrich, and Yinchu Zhu. A t-test for synthetic controls. *arXiv preprint arXiv:1812.10820*, 2018.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021a.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021b.

Stephen R Cosslett. Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica: Journal of the Econometric Society*, pages 765–782, 1983.

Yanqin Fan, Fang Han, Wei Li, and Xiao-Hua Zhou. On rank estimators in increasing dimensions. *Journal of Econometrics*, 214(2):379–412, 2020.

Bruno Ferman. On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116(536):1764–1772, 2021.

Bruno Ferman and Cristine Pinto. Placebo tests for synthetic controls. 2017.

Bruno Ferman and Cristine Pinto. Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221, 2021.

Sergio Firpo and Vitor Possebom. Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2), 2018.

Jean-Jacques Forneron. Estimation and inference by stochastic optimization. *arXiv preprint arXiv:2205.03254*, 2022.

Hiroshi Fujiki and Cheng Hsiao. Disentangling the effects of multiple treatments–measuring the net economic impact of the 1995 great hanshin-awaji earthquake. *Journal of Econometrics*, 186(1): 66–73, 2015.

Sebastian Galiani and Brian Quistorff. The synth_runner package: Utilities to automate synthetic control estimation using synth. *The Stata Journal*, 17(4):834–849, 2017.

Li Gan, Zhichao Yin, Nan Jia, Shu Xu, Shuang Ma, Lu Zheng, et al. Data you need to know about china. *Springer Berlin Heidelberg. https://doi*, 10:978–3, 2014.

Javier Gardeazabal and Ainhoa Vega-Bayo. An empirical comparison between the synthetic control method and hsiao et al.'s panel data approach to program evaluation. *Journal of Applied Econometrics*, 32(5):983–1002, 2017.

Jinyong Hahn and Ruoyao Shi. Synthetic control and inference. *Econometrics*, 5(4):52, 2017.

Peter Hall, Jeff Racine, and Qi Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

Aaron K Han. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3):303–316, 1987.

Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

Xuming He, Lan Wang, and Hyokyoung Grace Hong. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41(1):342–369, 2013.

Elhanan Helpman, Marc Melitz, and Yona Rubinstein. Estimating trade flows: Trading partners and trading volumes. *The quarterly journal of economics*, 123(2):441–487, 2008.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531, 1992.

Joel L Horowitz and Wolfgang Härdle. Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640, 1996.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

Cheng Hsiao and Qiankun Zhou. Panel parametric, semiparametric, and nonparametric construction of counterfactuals. *Journal of Applied Econometrics*, 34(4):463–481, 2019.

Cheng Hsiao, H Steve Ching, and Shui Ki Wan. A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740, 2012.

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

Shakeeb Khan, Xiaoying Lan, and Elie Tamer. Estimating high dimensional monotone index models by iterative convex optimization1. *arXiv preprint arXiv:2110.04388*, 2023.

Sungjin Kim, Clarence Lee, and Sachin Gupta. Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852, 2020.

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Arthur Lewbel. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of econometrics*, 97(1):145–177, 2000.

Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083, 2020.

Eckhard Liebscher. Strong convergence of sums of α-mixing random variables with applications to density estimation. *Stochastic Processes and Their Applications*, 65(1):69–80, 1996.

Charles F Manski. Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228, 1975.

Charles F Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of econometrics*, 27(3):313–333, 1985.

Nicolai Meinshausen. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

Katta G Murty and Santosh N Kabadi. Some np-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

W.K. Newey and F. Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.

J François Outreville. The relationship between relative risk aversion and the level of education: A survey and implications for the demand for life insurance. *Journal of economic surveys*, 29(1): 97–111, 2015.

Fu Ouyang and Thomas Tao Yang. High dimensional binary choice model with unknown heteroskedasticity or instrumental variables. 2023.

Xun Pang, Licheng Liu, and Yiqing Xu. A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288, 2022.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

James L Powell, James H Stock, and Thomas M Stoker. Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430, 1989.

Christoph Rothe. Semiparametric estimation of binary response models with endogenous regressors. *Journal of Econometrics*, 153(1):51–64, 2009.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Robert P Sherman. The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, pages 123–137, 1993.

Zhentao Shi and Jingyi Huang. Forward-selected panel data approach for program evaluation. *Journal of Econometrics*, 2021.

Y. Shin and Z. Todorov. Exact computation of the maximum rank correlation estimator. *Forthcoming, Econometrics Journal*, 2021.

Thomas M Stoker. Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481, 1986.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *The Annals of Statistics*, 45(4):1694–1727, 2017.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2016.

Shui-Ki Wan, Yimeng Xie, and Cheng Hsiao. Panel data approach vs synthetic control method. *Economics Letters*, 164:121–123, 2018.

HaiYing Wang. More efficient estimation for logistic regression with optimal subsamples. *Journal of machine learning research*, 20, 2019.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

Haiyong Wang and Shuhuang Xiang. On the convergence rates of legendre approximation. *Mathematics of computation*, 81(278):861–877, 2012.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

Kenneth Q Zhou and Stephen L Portnoy. Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics*, 24(1):287–306, 1996.