Estimating High Dimensional Monotone Index Models by Iterative Convex Optimization*

Shakeeb Khan[1], Xiaoying Lan[1], Elie Tamer[2], and Qingsong Yao[1]

[1]Dept. of Economics, Boston College

{ shakeeb.khan Xiaoyang.lan qingsong.yao}@bc.edu

[2]Dept. of Economics, Harvard University

{elietamer}@fas.harvard.edu

First Version: 6/2021, This Version: March 7, 2023

Abstract

We propose new approaches to estimating large dimensional semiparametric 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 our 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 optimize with more than a few regressors. Some recent progress is the work by Ahn, Ichimura, Powell, and Ruud (2018), but it too is only suitable for small dimensional models. Thus for such monotone index models with large, or even increasing dimension, we propose a new class of semiparametric sieve and kernel based estimators based on *batched gradient descent (BGD)*, 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.

Key Words Monotone Index models, Convex Optimization, Kernel and Sieve Estimation.

*We are grateful to conference participants at the BC/BU 2020 Econometrics Workshop, the 2019 Midwestern Econometrics Study group, the 2021 NASM of the Econometric Society, 2022 CIRAQ Econometrics conference, 2022 ISNPS conference, 2022 Advanced Methods Conference at TSE, and seminar participants from Georgetown, UC Berkeley, UC Louvain, UC Riverside, University of Bristol, UVA, University of Warwick and Yale for helpful comments.

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:

$$y_i = I[\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e^\star - u_i \geq 0] \tag{1}$$

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

[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.

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 survey 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. (2018).

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, where it is shown that χ^2 asymptotic approximations to 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 computation. 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. Finally, our proof in the semiparametric case requires development of approaches to handle estimators that are defined recursively while at the same time allow for an unknown link function. The paper starts out analyzing properties of the BGD estimator in parametric models with increasing dimensions. The following section considers the main case when we allow for the link function to be estimated via kernel or sieve methods. Finally a Monte Carlo Section examines the computational advantages of our approach. Also, we provide an empirical illustration that highlights the behavior of our estimator with real data.

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

Notation: 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 $\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$. 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 $\|f\|_{\infty} = \sup_{\boldsymbol{x}\in D} f(\boldsymbol{x})$.

2 The BGD Estimator

To provide some intuition for our semiparametric estimators that will be introduced in the following sections, we first consider here 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. *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

[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.

respect to $\boldsymbol{\beta}_e \in \mathcal{B}_e$, the BGD estimator of $\boldsymbol{\beta}_e^\star$ is based on the following iteration,

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \partial \ell_G \left(\boldsymbol{\beta}_{e,k}, \mathbf{X}_{e,i}, y_i \right) / \partial \boldsymbol{\beta}_e, \tag{2}$$

where $\delta_k > 0$ is the learning rate. Note that $n^{-1} \sum_{i=1}^{n} \partial \ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_{e,i}, y_i \right) / \partial \boldsymbol{\beta}_e$ constitutes a sample analogue of the derivative $\partial \mathbb{E} \left[\ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right) \right] / \partial \boldsymbol{\beta}_e$. 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 (2) until some terminating conditions are reached.

In this paper, we consider the following loss function

$$\ell_G \left(\boldsymbol{\beta}_e, \mathbf{X}_e, y \right) = \int_{-A}^{\mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e} G(z)\, dz - y \mathbf{X}_e^{\mathrm{T}} \boldsymbol{\beta}_e, \tag{3}$$

for some sufficiently large positive constant A. The loss function (3) was also considered in Agarwal et al. (2014) and has many properties. For instance, under some mild conditions, it is easy to show that at the truth,

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

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 (3) is that the derivative of (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 $G(\cdot)$ instead of on its derivatives. So when we conduct a semiparametric iteration in the following sections, we only need to nonparametrically approximate $G(\cdot)$, which is generally more robust compared with approximating its derivatives. Based on loss function (3), the BGD estimator is obtained by using the following iteration procedure:

$$\boldsymbol{\beta}_{e,k+1} = \boldsymbol{\beta}_{e,k} - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(G \left(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_{e,k} \right) - y_i \right) \mathbf{X}_{e,i}. \tag{4}$$

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 $G(\cdot)$, and terminating condition $\mathcal{T}$

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

$k \leftarrow 1$;

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

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

  $k \leftarrow k + 1$;

$\widehat{\boldsymbol{\beta}}_e \leftarrow \boldsymbol{\beta}_{e,k}$;

Remark 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 $y_i = G(\mathbf{X}_{e,i}^{\mathrm{T}} \boldsymbol{\beta}_e) + \varepsilon_i$ with $E[\varepsilon_i | \mathbf{X}_{e,i}] = 0$ and monotone $G(.)$, a similar *convex* criterion as in (3) can be used for inference on $\boldsymbol{\beta}_e$.

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

Assumption 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* $\boldsymbol{\beta}_e \in \mathcal{B}_e$, $|\beta_j| \le B_0$ *for any* $0 \le j \le p$*; (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* $0 < \underline{\lambda}_e \le \underline{\lambda}(M(\boldsymbol{\beta}_e)) \le \overline{\lambda}(M(\boldsymbol{\beta}_e)) \le \overline{\lambda}_e < \infty$*.*

Remark 2. Assumption 2(i) and Assumption 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 (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 2(iii) imposes some smoothness conditions on G, where the requirement on υ_G will be stated in the following propositions and theorems. Assumption 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 $\mathbb{E}\left[\varepsilon_i | \mathbf{X}_{e,i} \right] = 0$. When Assumption 1 and Assumption 2 hold, we have the following result.

Theorem 1. *Suppose that Assumption 1 and Assumption 2 hold with $\upsilon_G = 3$, that $p^5 (\log p)^2 n^{-1} \to 0$, that the learning rate is chosen such that $\delta_k = \delta \le 2/\left(3\overline{\lambda}_e\right)$, 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

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

(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

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

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

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

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. See Appendix B. □

When p is fixed, Theorem 1(i) implies that $\sup_{k \ge k_{1,n}^{BGD}+1}\left\|\Delta\boldsymbol{\beta}_{e,k}\right\| = O_p\left(1/\sqrt{n}\right)$, and Theorem 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(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 $O\left(\log n\right)$, 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 5 at the end of Section 4. The inference on $\boldsymbol{\beta}_e^\star$ based on the BGD estimator is given by Theorem 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. Different from the stochastic gradient descent algorithm (e.g., Toulis and Airoldi, 2017), we show in Theorem 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. In particular, we have the following proposition.

Theorem 2. *Suppose that all the conditions in Theorem 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(iii) and (iv) hold.

Proof of Theorem 2. See Appendix B. □

Theorem 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, 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(i); when $\upsilon > 0$, we can see that more rounds of iteration is needed compared with required in Theorem 1(i).

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 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 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 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 $G(z) = \mathbb{E}\left[y \mid X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^\star = z\right] = \mathbb{E}\left[y \mid X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_k = z\right]$ 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.

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

$$\widehat{G}\left(z \middle| \boldsymbol{\beta}_k\right) = \frac{\sum_{j=1}^{n} K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}} \boldsymbol{\beta}_k\right) y_j}{\sum_{j=1}^{n} K_{h_n}\left(z - X_{0,j} - \mathbf{X}_j^{\mathrm{T}} \boldsymbol{\beta}_k\right)}, z \in R, \tag{5}$$

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

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n} \sum_{i=1}^{n} \left(\widehat{G}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k \middle| \boldsymbol{\beta}_k \right) - y_i \right) \mathbf{X}_i. \tag{6}$$

Keep updating $\boldsymbol{\beta}_k$ based on (5) and (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 ← 1;
while The terminating condition T is not satisfied do
    for i ← 1 to n do
        Ĝ(X_{0,i} + X_i^T β_k | β_k) ← Σ_{j=1}^n K_{h_n}(X_{0,i} + X_i^T β_k − X_{0,j} − X_j^T β_k) y_j / Σ_{j=1}^n K_{h_n}(X_{0,i} + X_i^T β_k − X_{0,j} − X_j^T β_k);
    β_{k+1} ← β_k − (δ_k / n) Σ_{i=1}^n (Ĝ(X_{0,i} + X_i^T β_k | β_k) − y_i) X_{e,i};
    k ← k + 1;
β̂ ← β_k;
```

Remark 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 (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 (3) 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 (6). On the contrary, during the construction (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

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

so additional terms involving $\partial \widehat{G}(\cdot|\,\boldsymbol{\beta}_k)/\partial \boldsymbol{\beta}$ are avoided. Finally, as we discussed in Section 2, the derivative of loss function (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 Assumption 5.

For any fixed z and $\boldsymbol{\beta}$, under mild conditions there holds $\widehat{G}(z|\,\boldsymbol{\beta}) \to_p \mathbb{E}\left[y|\, X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} = z\right]$. Denote such limit as $L(z, \boldsymbol{\beta})$. Obviously, $L(z, \boldsymbol{\beta}^{\star}) = G(z)$ holds for any $z \in \mathbb{R}$. Before we move to a formal description of the statistical properties of the KBGD estimator based on (6), we first provide some further discussion on $L(z, \boldsymbol{\beta})$. 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 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 $f_e(\mathbf{X}_e) = f_e(X_0, \mathbf{X})$ and $f(\mathbf{X}) = \int f_e(X_0, \mathbf{X})\, dX_0$, respectively. Denote $z(\mathbf{X}_e, \boldsymbol{\beta}) = X_0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. Also denote $f_{\mathbf{X},z}(\mathbf{X}, z|\,\boldsymbol{\beta})$ as the joint density of $\mathbf{X}$ and $z(\mathbf{X}_e, \boldsymbol{\beta})$ given $\boldsymbol{\beta}$. Note that for any $\boldsymbol{x}$

and z,

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

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

$$
f_{\mathbf{X},z}\left(\mathbf{X}, z \middle| \boldsymbol{\beta}\right) = f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right), \tag{7}
$$

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

$$
f_z\left(z \middle| \boldsymbol{\beta}\right) = \int_{\mathcal{X}} f_{\mathbf{X},z}\left(\mathbf{X}, z \middle| \boldsymbol{\beta}\right) d\mathbf{X} = \int_{\mathcal{X}} f_e\left(z - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{X}\right) d\mathbf{X}. \tag{8}
$$

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

$$
\begin{aligned}
L\left(z, \boldsymbol{\beta}\right) &= \mathbb{E}\left(G\left(z - \mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right) \middle| z\left(\mathbf{X}_e, \boldsymbol{\beta}\right) = z\right) \\
&= \int_{\mathcal{X}} G\left(z - \mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right) f_{\mathbf{X}|z}\left(\mathbf{X} \middle| z, \boldsymbol{\beta}\right) d\mathbf{X},
\end{aligned} \tag{9}
$$

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 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 $1 \leq \upsilon \leq \upsilon_K - 1$ and $\int u^{\upsilon_K} K(u)\,du \neq 0$; (iv) $K(s) = 0$ for $|s| > 1$.*

Assumption 4. *(i) There exists some constant $\zeta > 1$ such that $\zeta^{-1} \leq f_e\left(\mathbf{X}_e\right) \leq \zeta$ holds for all $\mathbf{X}_e \in \mathcal{X}_e$; (ii) there exists positive integer υ_f such that $f_e\left(\mathbf{X}_e\right)$ has bounded up to υ_f-th derivatives.*

Remark 4. Assumption 4(i) together with Assumption 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 4(i) basically allows us to construct a subset of $\mathcal{X}_e$ such that $f_z\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)|\boldsymbol{\beta}\right)$ is uniformly lowered bounded from zero over such subset.

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

Lemma 1. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 3, and Assumption 4 hold with* $\upsilon_G=3$, $\upsilon_K=2$, *and* $\upsilon_f=3$. *Define* $\psi\left(n,p,h\right)=h^{-1}\sqrt{\log\left(pnh^{-1}\right)/n}+h^2$. *If* $h_n\to 0$ *and* $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\to 0$ *further hold, we have that*

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

Proof of Lemma 1. See Appendix A. □

Lemma 1 implies that $\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(\left.Z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right|\boldsymbol{\beta}\right)\mathbf{X}_i$ will be closer to $\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]$ 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 $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)$ is small. So even when $\widehat{G}\left(\left.z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\right|\boldsymbol{\beta}\right)$ is a poor estimator for $L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)$ 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, the dimension p must satisfy $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right)\to 0$. Suppose that $p/n\to 0$ and we choose $h_n=\left(\left(\log n\right)/n\right)^{1/6}$, we have that $\psi\left(n,p,h_n\right)\sim\left(\left(\log n\right)/n\right)^{1/3}$. This implies that when p is fixed, the convergence rate in Lemma 1 is $\left(\left(\log n\right)/n\right)^{1/3(p+1)}$. When p increases with n, the dimension p should satisfy $p\log p=O\left(\log n\right)$, 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 $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)$ 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 $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_k\right)$ and $z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}^\star\right)$ as $z_{i,k}$ and $z_i^\star$. Based on the results in Lemma 1, we have that under all conditions as imposed in Lemma 1, there holds

$$\boldsymbol{\beta}_{k+1}=\boldsymbol{\beta}_k-\delta_k\mathbb{E}\left[\left(L\left(z_{i,k},\boldsymbol{\beta}_k\right)-G\left(z_i^\star\right)\right)\cdot\mathbf{X}_i\right]+\delta_k\cdot\left(\text{small order terms}\right). \tag{10}$$

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

$(L(z_{i,k}, \boldsymbol{\beta}_k) - G(z_i^\star)) \cdot \mathbf{X}_i$ equals to

$$\left\{ \int_{\mathcal{X}} \left[G\left(z_i^\star + \mathbf{X}_i^{\mathrm{T}} \Delta\boldsymbol{\beta}_k - \mathbf{X}^{\mathrm{T}} \Delta\boldsymbol{\beta}_k \right) - G(z_i^\star) \right] f_{\mathbf{X}|z}(\mathbf{X}|\, z_{i,k}, \boldsymbol{\beta}_k)\, d\mathbf{X} \right\} \cdot \mathbf{X}_i$$
$$= \int_0^1 \int_{\mathcal{X}} \left[G'\left(z_i^\star + t\left(\mathbf{X}_i - \mathbf{X}\right)^{\mathrm{T}} \Delta\boldsymbol{\beta}_k \right) f_{\mathbf{X}|z}(\mathbf{X}|\, z_{i,k}, \boldsymbol{\beta}_k) \left(\mathbf{X}_i \mathbf{X}_i^{\mathrm{T}} - \mathbf{X}_i \mathbf{X}^{\mathrm{T}} \right) \right] \Delta\boldsymbol{\beta}_k d\mathbf{X} dt, \quad (11)$$

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

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

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

and

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

we have that

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

which indicates that

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

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

Assumption 5. *There hold*

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

and

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

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

Theorem 3. *Suppose that Assumption 1, Assumption 2(i)–(iii), Assumption 3–Assumption 5 hold with* $\upsilon_G = 3$, $\upsilon_K = 2$, *and* $\upsilon_f = 3$, $\delta_k = \delta$ *such that* $\delta < \min\left\{ 1/\left(2\underline{\lambda}_{\mathit{\Lambda}}\right), 1/\left(4p^2 \|G'\|_\infty\right) \right\}$, *and that $\boldsymbol{\beta}$ is updated based on algorithm 2. Define*

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

Then if $h_n \to 0$ and $p^{\frac{5p+1}{2(p+1)}}\psi^{\frac{1}{p+1}}\left(n,p,h_n\right) \to 0$ hold, we have that

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

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

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

Proof of Theorem 3. See Appendix B. □

Theorem 3 implies that the iterative estimator based on (5) and (6) is consistent under increasing 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 as we have discussed in Lemma 7.

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 3, even in the case of fixed dimensionality, the guaranteed convergence rate of the KBGD estimator based on (5) and (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,

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n} I_i^{\phi}\cdot\left(\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) - y_i\right)\mathbf{X}_i, \tag{12}$$

where $\widehat{G}\left(z_{i,k}\middle|\boldsymbol{\beta}_k\right) = \widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_k\right)\middle|\boldsymbol{\beta}_k\right)$ is defined in (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

$$\mathcal{X}_e^{\phi} = \{\mathbf{X}_e \in \mathcal{X}_e : |X_j| \leq 1 - \phi, 0 \leq j \leq p\} \tag{13}$$

for some $\phi > 0$ whose value will be determined later. Different from (6), the update of $\boldsymbol{\beta}_k$ based on (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 (13) is that, as we show in the Appendix A, for any $0 < \phi < 1$, there holds $\inf_{(\mathbf{X}_e, \boldsymbol{\beta}) \in \mathcal{X}_e^{\phi} \times \mathcal{B}} f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right) \middle| \boldsymbol{\beta}\right) \geq C\phi^p p^{-p}$ for some constant $C > 0$ that depends on ϕ. When p and ϕ are both fixed, $f_z\left(z\left(\mathbf{X}_e, \boldsymbol{\beta}\right) \middle| \boldsymbol{\beta}\right)$ 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

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

The following theorem provides a counterpart to the results in Theorem 3.

Theorem 4. *Suppose that all the assumptions and conditions on υ_G, υ_K, and υ_f in Theorem 3 hold. Suppose moreover that $h_n \to 0$, $\delta_k = \delta < \min\left\{1/\left(2\underline{\lambda}_{\Lambda}\right), 1/\left(4p^2 \|G'\|_{\infty}\right)\right\}$, $\phi < \delta\underline{\lambda}_{\Lambda}/\left(16p^2 \|G'\|_{\infty} \zeta\right)$, and that $\boldsymbol{\beta}$ is updated under (5) and (12) (The trimmed version of algorithm 2). Define*

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

then there holds

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

Proof of Theorem 4. See Appendix B. □

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 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

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

We note that

$$\begin{aligned}
\Delta\boldsymbol{\beta}_{k+1} &= \Delta\boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\right|\boldsymbol{\beta}_k) - y_i\right)\mathbf{X}_i^{\phi}, \\
&= \Delta\boldsymbol{\beta}_k - \frac{\delta_k}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z_{i,k}\right|\boldsymbol{\beta}_k) - \widehat{G}\left(z_i^{\star}\right|\boldsymbol{\beta}^{\star})\right)\mathbf{X}_i^{\phi} - \delta_k\boldsymbol{\xi}_n^{\phi} \\
&= \int_0^1\left\{I_p - \frac{\delta_k}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^{\phi}\left.\frac{\partial\widehat{G}\left(z\left(\boldsymbol{X}_{e,i},\boldsymbol{\beta}\right)\right|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right]\right\}dt\Delta\boldsymbol{\beta}_k - \delta_k\boldsymbol{\xi}_n^{\phi},
\end{aligned} \tag{14}$$

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

Lemma 2. *Suppose that all the assumptions in Theorem 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*

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

Proof of Lemma 2. See Appendix A. ☐

Lemma 3. *Suppose that all the assumptions in Theorem 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 3. See Appendix A. ☐

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

Theorem 5. *Suppose that all the assumptions in Theorem 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 $h_n^4 n/(\log n)^2 \to \infty$, and that $\boldsymbol{\beta}$ is updated under (5) and (12). Then*

(i) There holds

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

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 Theorem 5. See Appendix B. □

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

Theorem 6. *Suppose that all the assumptions and conditions in Theorem 5 hold. Suppose also that $\widehat{\boldsymbol{\beta}}$ is defined as in Theorem 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

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

Proof of Theorem 6. See Appendix B. □

We finally provide some remarks for the KBGD estimators.

Remark 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 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, Theorem 5 requires that h_n is chosen such that $nh_n^6 \to 0$ and $nh_n^4/\left(\log n\right)^2 \to \infty$. 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 $C = C_k = \operatorname{std}\left(z_{i,k}\right)$ for the k-th round of iteration and $C = \operatorname{std}\left(\widehat{z}_i\right)$ when estimating the variance $\Sigma_{\boldsymbol{\beta}}^{\phi}$. We finally discuss the terminating conditions. As we show in Theorem 5, to obtain root-n consistency and asymptotic normality, the iteration number is required to be only of order $\log\left(n\right)$. However, such rule can not be directly applied to determine the number of iterations since the initial distance $\left\|\Delta\boldsymbol{\beta}_1\right\|$ 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 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 $f_{\mathbf{X}_c|\mathbf{X}_d}\left(\mathbf{X}_c|\mathbf{X}_d\right)$. Then we require that all the conditions imposed on the $f_e\left(\mathbf{X}_e\right)$ hold for $f_{\mathbf{X}_c|\mathbf{X}_d}\left(\mathbf{X}_c|\mathbf{X}_d\right)$ for any realizations of $\mathbf{X}_d$.

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 $\{r_j\left(z\right)\}_{j=0}^{\infty}$ that is complete in $C\left(\mathbb{R}\right)$ space, any smooth CDF G can be represented by $G\left(z\right) = \sum_{j=0}^{\infty} \pi_j^{\star} r_j\left(z\right)$ 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 $\boldsymbol{r}_q(z) = \left(r_0\left(z\right), \cdots, r_q\left(z\right)\right)^{\mathrm{T}}$ and $\boldsymbol{\pi}_q^{\star} = \left(\pi_1^{\star}, \cdots, \pi_q^{\star}\right)^{\mathrm{T}}$, we have that

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

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

$$\widehat{\boldsymbol{\pi}}_{q,n,k} = \left(\sum_{i=1}^{n} \boldsymbol{r}_q\left(z_{i,k}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{r}_q\left(z_{i,k}\right) y_i\right). \tag{15}$$

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

$$\widehat{G}\left(z\middle|\boldsymbol{\beta}_k\right)=\boldsymbol{r}_q^{\mathrm{T}}\left(z\right)\widehat{\boldsymbol{\pi}}_{n,q,k},\ -\infty<z<\infty. \tag{16}$$

Based on the estimated CDF $\widehat{G}\left(z\middle|\boldsymbol{\beta}_k\right)$, the update of the parameter can be carried out based on (6). We iterate sequentially based on (15), (16) and (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.

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 $\boldsymbol{r}(z)=r_0\left(z\right),\cdots,r_q\left(z\right)$, 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**

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

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

  $k\leftarrow k+1$;

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

Remark 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 $\widehat{G}\left(z\middle|\boldsymbol{\beta}_k\right)$ 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 8. Compared with the KBGD algorithm, the SBGD procedure has at least two advantages. 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 (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 $R_q\left(z\right)=G\left(z\right)-\boldsymbol{r}^{\mathrm{T}}\left(z\right)\boldsymbol{\pi}_q^{\star}$, $\Gamma_{q,n}\left(\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, $\Gamma_{q,n,k}=\Gamma_{q,n}\left(\boldsymbol{\beta}_k\right)$, and $\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i}+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_{q,n}^{-1}\left(\boldsymbol{\beta}\right)\boldsymbol{r}_q\left(z\right)\mathbf{X}_i\right)$. Through tedious algebra, we can show that the SBGD procedure has the following representation,

$$\begin{aligned}\boldsymbol{\beta}_{k+1}&=\boldsymbol{\beta}_k-\frac{\delta_k}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i-\mathfrak{X}_{q,n}\left(z_{i,k},\boldsymbol{\beta}_k\right)\right)\left(G\left(z_{i,k}\right)-G\left(z_i^{\star}\right)\right)\\&\quad-\frac{\delta_k}{n}\sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{r}_q^{\mathrm{T}}\left(z_{i,k}\right)\Gamma_{q,n,k}^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)R_q\left(z_{j,k}\right)+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{r}_q\left(z_{j,k}\right)\varepsilon_j\right)\\&\quad+\frac{\delta_k}{n}\sum_{i=1}^{n}\left(R_q\left(z_{i,k}\right)\mathbf{X}_i+\varepsilon_i\mathbf{X}_i\right),\end{aligned}\tag{17}$$

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 6. *(i) There holds* $\max_{0\le j\le q}\left\|r_j\right\|_{\infty}\le D_{q,0}$, $\max_{0\le j\le q}\left\|r_j'\right\|_{\infty}\le D_{q,1}$, *and* $\max_{0\le j\le q}\left\|r_j''\right\|_{\infty}\le D_{q,2}$; *(ii) Define* $\Gamma_q\left(\boldsymbol{\beta}\right)=\mathbb{E}\left(\boldsymbol{r}_q\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{r}_q^{\mathrm{T}}\left(X_0+\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$, *there hold* $\inf_{\boldsymbol{\beta}\in\mathcal{B}}\underline{\lambda}\left(\Gamma_q\left(\boldsymbol{\beta}\right)\right)\ge\underline{\lambda}_{\Gamma}>0$ *and* $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\lambda}\left(\Gamma_q\left(\boldsymbol{\beta}\right)\right)\le\overline{\lambda}_{\Gamma}<\infty$ *for all* q; *(iii) There hold* $\sup_{z\in R}\left|G\left(z\right)-\boldsymbol{r}^{\mathrm{T}}\left(z\right)\boldsymbol{\pi}_q^{\star}\right|\le\mathcal{E}_{q,0}$ *and* $\sup_{z\in R}\left|G'\left(z\right)-\left(\boldsymbol{r}'\left(z\right)\right)^{\mathrm{T}}\boldsymbol{\pi}_q^{\star}\right|\le\mathcal{E}_{q,1}$, *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 $\mathfrak{X}_{q,n}\left(z,\boldsymbol{\beta}\right)$ as

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

Then we have the following lemma.

Lemma 4. *Define* $\chi_{1,n}=\sqrt{pq^2D_{q,0}^4\log\left(pqD_{q,0}D_{q,1}n\right)/n}$, *and* $\chi_{2,n}=\sqrt{p}qD_{q,0}^2\left(\chi_{1,n}+\mathcal{E}_{q,0}\right)$. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold, and moreover,* $\upsilon_G\ge 1$ *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 $\sup_{k\geq 1}\|\mathfrak{R}_{n,k}\|=O_p\left(\chi_{2,n}\right)$.

Proof of Lemma 4. See Appendix A. □

Obviously, Lemma 4 provides a parallel result to (10). In particular, define

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

under all the conditions imposed in Lemma 4, we have that

$$\Delta\boldsymbol{\beta}_{k+1}=\left\{\int_0^1\left(I_p-\delta_k\Psi_q\left(t,\boldsymbol{\beta}_k\right)\right)dt\right\}\Delta\boldsymbol{\beta}_k+\delta_k\mathfrak{R}_{n,k}. \tag{18}$$

Obviously, (18) is also a parallel result to (11). As a result, to ensure that (18) actually constitutes a contraction for $\|\Delta\boldsymbol{\beta}_k\|$, we impose the following assumption that is similar to Assumption 5.

Assumption 7. *For any $q\geq 0$, there hold*

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

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

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

Theorem 7. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 6 and Assumption 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 $0<\delta<\min\left\{1/\left(2\underline{\lambda}_\Psi\right),\underline{\lambda}_\Psi/\left(2\|G'\|_\infty^2p^2\left\{1+\underline{\lambda}_\Gamma^{-1}qD_{q,0}^2\right\}^2\right)\right\}$, and that $\boldsymbol{\beta}$ is updated based on algorithm 3. Define*

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

then we have that

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

Proof of Theorem 7. See Appendix B. □

According to Theorem 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 (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 2 and Lemma 3, we provide two additional lemmas that are useful to understand the above algorithm.

Lemma 5. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 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*

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

Proof of Lemma 5. See Appendix A. □

Lemma 6. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 6, and Assumption 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*

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

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 6. See Appendix A. □

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

Theorem 8. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 6 and Assumption 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 υ_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 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho^{\mathrm{T}}\Psi_q^{\star-1}\left(\mathbf{X}_i-\mathfrak{X}_q\left(z_i^\star,\boldsymbol{\beta}^\star\right)\right)\varepsilon_i\to_d$

$N\left(0, \sigma_S^2(\rho)\right)$ *with*

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

there holds

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

.

Proof of Theorem 8. See Appendix B. □

We now provide the estimator for the variance.

Theorem 9. *Suppose that all the conditions listed in Theorem 8 hold and* $pq^2 D_{q,0}^4 \mathcal{E}_{q,1} \to 0$ *as* $n \to 0$*. Let* $\widehat{\boldsymbol{\beta}}$ *be as defined as in Theorem 8. Define* $\widehat{\boldsymbol{r}}_{q,i} = \boldsymbol{r}_q\left(z\left(\mathbf{X}_{e,i}, \widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{r}}'_{q,i} = \boldsymbol{r}'_q\left(z\left(\mathbf{X}_{e,i}, \widehat{\boldsymbol{\beta}}\right)\right)$, $\widehat{\boldsymbol{\pi}}_q = \left(\sum_{i=1}^n \widehat{\boldsymbol{r}}_{q,i}\widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^n \widehat{\boldsymbol{r}}_{q,i} y_i\right)$, $\widehat{G}_i = \widehat{\boldsymbol{r}}_{q,i}^{\mathrm{T}}\widehat{\boldsymbol{\pi}}$, $\widehat{G}'_i = \widehat{\boldsymbol{r}}'^{\mathrm{T}}_{q,i}\widehat{\boldsymbol{\pi}}_q$, $\widehat{\Psi}^\star_{q,i} = \frac{1}{n}\sum_{i=1}^n \widehat{G}'_i \cdot \left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} - \mathfrak{X}_{q,n}\left(\widehat{z}_i, \widehat{\boldsymbol{\beta}}\right)\mathbf{X}_i^{\mathrm{T}}\right)$, $\widehat{\mathfrak{X}}_{q,i} = \frac{1}{n}\sum_{j=1}^n \mathbf{X}_j \widehat{\boldsymbol{r}}_{q,j}^{\mathrm{T}} \Gamma_{q,n}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)\widehat{\boldsymbol{r}}_{q,i}$, *and*

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

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 9. See Appendix B. □

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

Remark 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 $\{1, z, z^2, \cdots, z^q\}$, the correlation between the sieve functions increases as the approximation order q increases, which may lead to a violation of Assumption 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 5, we use $T(z) = 2\pi^{-1}\arctan(z)$. For the uniform error bound of truncated Legendre polynomials, see Wang and Xiang (2012).

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 $\text{Bias} = |\frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_j^r - \beta_j^\star)|$ and $\text{RMSE} = \sqrt{\sum_{r=1}^{R}(\widehat{\beta}_j^r - \beta_j^\star)^2/R}$. 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. 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 $0 \leq j \leq 8$, $X_{9,i} \sim \text{Bernoulli}(1/2)$, $X_{10,i} \sim \text{Poisson}(2)$, and $u_i \sim \textit{Cauchy}$. 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 reports the finite-sample properties of our estimators. It can be seen that our estimators 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: 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 $h_n = \sigma_{\hat{z}} \cdot n^{-1/5}$, where n is sample size, $\sigma_{\hat{z}}$ 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 9. For both estimators, the stopping rule is either $\max_{1 \leq j \leq p} |\widehat{\beta}_{j,k+1} - \widehat{\beta}_{j,k}| < 10^{-5}$ or $k \geq 20000$. The above also applies to our empirical analysis in Section 5. 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.

Table 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.

Table 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

Table 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 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.

5 Empirical Illustration

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. We have a total of 11 explanatory variables in our model. The key factor that we are particularly interested in is the educational backgrounds, which is defined 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, and total asset, whose impacts on the risk aversion are of interest on their own right. When conducting semiparametric estimation, we normalize the coefficient of total asset to 1. 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 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 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 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. 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.

Table 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 $C = C_k = \mathrm{std}(z_{i,k})$, and use the fourth-order Epanechinikov kernel. For SBGD estimator, we choose $q = 9$ and use Legendre polynomials with transformation discussed in Remark 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.

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.

A 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 A.1. *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, $\sup_{U \in \Omega, \theta \in \Theta} \|A_{s,t}(U, \theta)\| \leq D_{A,0}$ and $\sup_{U \in \Omega} \|A_{s,t}(U, \theta_1) - A_{s,t}(U, \theta_2)\| \leq D_{A,1} \|\theta_1 - \theta_2\|$ uniformly for all $1 \leq s \leq d_1$ and $1 \leq t \leq d_2$. Then there holds*

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n A(U_i, \theta) - \mathbb{E}A(U_i, \theta) \right\| = O_p\left(\sqrt{\frac{p d_1 d_2 D_{A,0}^2 \log(d_1 d_2 D_{A,1} n)}{n}} \right).$$

Proof of Lemma A.1. Note that

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

For the first term, we have that

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

indicating that

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

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

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

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

Lemma A.2. *If Assumption 1, Assumption 2(i)-(iii), and Assumption 4 hold with* $\min\{\upsilon_G,\upsilon_f\}\ge 2$*, then there exists a constant* C *that does not depend on* $\mathbf{X},z,\boldsymbol{\beta}$ *such that the following hold*

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

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

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

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

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

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

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

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

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

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

Proof. To prove Lemma A.2(i) and Lemma A.2(ii), we note that for any $0\le s\le\upsilon_f$,

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

and

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

Since $f_e\left(\mathbf{X}_e\right)$ has up to υ_f-th bounded derivatives over $\mathcal{X}_e$ according to Assumption 4(ii) and X_j is bounded by 1 for all $1\le j\le p$ according to Assumption 2(i), Lemma A.2(i) and Lemma A.2(ii) hold.

Similarly, note that

$$\frac{\partial f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}=-\left[\left.\frac{\partial f_{e}\left(X_{0},\mathbf{X}\right)}{\partial X_{0}}\right|_{X_{0}=z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right]\mathbf{X},$$

$$\frac{\partial^{2} f_{\mathbf{X},z}\left(\mathbf{X},z|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}=\left[\left.\frac{\partial^{2} f_{e}\left(X_{0},\mathbf{X}\right)}{\partial X_{0}^{2}}\right|_{X_{0}=z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right]\mathbf{X}\mathbf{X}^{\mathrm{T}},$$

$$\frac{\partial f_{z}\left(z|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}}=-\int_{\mathcal{X}}\left[\left.\frac{\partial f_{e}\left(X_{0},\mathbf{X}\right)}{\partial X_{0}}\right|_{X_{0}=z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right]\mathbf{X}d\mathbf{X},$$

$$\frac{\partial^{2} f_{z}\left(z|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}}=\int_{\mathcal{X}}\left[\left.\frac{\partial^{2} f_{e}\left(X_{0},\mathbf{X}\right)}{\partial X_{0}^{2}}\right|_{X_{0}=z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right]\mathbf{X}\mathbf{X}^{\mathrm{T}}d\mathbf{X},$$

we validate Lemma A.2(iii)-Lemma A.2(vi).

To prove Lemma A.2(vii), note that

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

According to Assumption 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 $\int_{\mathcal{X}}\left|\partial^{s-j} f_{\mathbf{X}|z}\left(\mathbf{X}|\,z,\boldsymbol{\beta}\right)/\partial z^{s-j}\right|d\mathbf{X}=1$. When $j=s-1$, define $\mathbb{X}\left(z,\boldsymbol{\beta}\right)=\left\{\mathbf{X}:\left(z-\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta},\mathbf{X}\right)\in\mathcal{X}_{e}\right\}$. We have that

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

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 A.2(viii), note that

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

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

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

according to part (iii) of this lemma. This proves Lemma A.2(viii). Lemma A.2(ix) can be similarly proved.

Finally, to show Lemma A.2(x), we note that

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

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

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

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 $\left\| \partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z \middle| \boldsymbol{\beta}\right)/\partial z\Big|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}\right\|$ is bounded, and $\left\| \partial f_{\mathbf{X},z}\left(\widetilde{\mathbf{X}}, z \middle| \boldsymbol{\beta}\right)/\partial \boldsymbol{\beta}\Big|_{z=z(\mathbf{X}_e,\boldsymbol{\beta})}\right\|$ 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

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

This finishes the proof of Lemma A.2(xii). □

Lemma A.3. *Suppose that Assumption 1, Assumption 2(i)-(iii), 3 and Assumption 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

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

(ii) There holds

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

(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

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

Proof. Lemma A.3(i) is a direct result of Lemma A.1 if we note that

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

and

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

To prove Lemma A.3(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}_jdz\\
&=\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 A.2 (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 A.3 (iii) is a combination of Lemma A.3 (i) and Lemma A.3 (ii). $\square$

Lemma A.4. *Suppose that Assumption 1, Assumption 2(i)-(iii), Assumption 3, and Assumption 4 hold. Given any positive sequence $\{\phi_n\}_{n=1}^{\infty}$ satisfying $p\phi_n\downarrow 0$, define*

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

Then

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

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

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

Proof. To prove Lemma A.4(i), note that for $p\phi_n<1$, $m\left(\mathcal{X}_e-\mathcal{X}_{e,n}\right)=1-\left(1-\phi_n\right)^p\le p\phi_n$. So $\int_{\mathcal{X}_e-\mathcal{X}_{e,n}} f_e\left(\mathbf{X}_e\right)d\mathbf{X}_e\le\zeta p\phi_n=O\left(p\phi_n\right)$ due to Assumption 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 $|z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)-\widetilde{\mathbf{X}}^{\mathrm{T}}\boldsymbol{\beta}-X_0|\le\sum_{j=1}^p|\beta_j|\,|X_j-\widetilde{X}_j|$. This implies that for any $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{X}}\in\mathbb{X}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)$ if

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

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

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

due to Assumption 4(i). This proves Lemma A.4(i).

To prove Lemma A.4(ii), note that for any $\mathbf{X}_e$ and $\boldsymbol{\beta}$, we have $\widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)=A_{n,y}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)/A_{n,1}\left(\mathbf{X}_e,\boldsymbol{\beta}\right)$ 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

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

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 A.2(vii). Then the results hold according to Lemma A.3. □

Proof of Lemma 1.

Proof. Note that

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

$$\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)\mathbf{X}_i\right\| \tag{1}$$

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

Obviously, (1) is bounded by

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\right)\mathbf{X}_i\right\|$$

$$\leq\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot I_{n,i} \tag{3}$$

$$+\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot\left(1-I_{n,i}\right), \tag{4}$$

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 A.4. Note that (3) is bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot I_{n,i}$$

$$\leq\sup_{(\mathbf{X}_e,\boldsymbol{\beta})\in\mathcal{X}_{e,n}\times\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}-L\left(Z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}\right\|$$

$$=O_p\left(p^{p+1/2}\phi_n^{-p}\psi_1\left(n,p,h_n\right)\right),$$

according to Lemma A.4. For (4), we have that

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\boldsymbol{\beta}\right)\mathbf{X}_i-L\left(Z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right\|\cdot\left(1-I_{n,i}\right)$$

$$\leq C\sqrt{p}\,\mathbb{E}I\left(\mathbf{X}_{e,i}\notin\mathcal{X}_{e,n}\right)=O\left(p^{3/2}\phi_n\right),$$

according to Lemma A.4(i). Then we have that (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 (2). Similar to the above truncation, we have that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\right]\right\|\\
&\leq\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot I_{n,i}-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot I_{n,i}\right]\right\| &(5)\\
&+\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot\left(1-I_{n,i}\right)-\mathbb{E}\left[L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i\cdot\left(1-I_{n,i}\right)\right]\right\|. &(6)
\end{aligned}$$

Obviously, (6) is $O_p\left(p^{3/2}\phi_n\right)$. For (5), note that $\|L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)X_{j,i}\cdot I_{n,i}\|$ is bounded by C and $\partial\|L\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)X_{j,i}\cdot I_{n,i}/\partial\boldsymbol{\beta}\|$ is bounded by $C\sqrt{p}$ by Lemma A.2(vii) and (viii), we have that (5) is of order $O_p\left(\sqrt{p^2 n\log(pn)/n}\right)$ using Lemma A.1. 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

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

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

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

This finishes the whole proof. □

Lemma A.5. *Suppose that p is fixed. If all the assumptions in Lemma A.3 hold with $\upsilon_G=4$, $\upsilon_K=3$, and $\upsilon_f=4$, we have that Lemma A.3(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 A.3 so is omitted. □

Lemma A.6. *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 A.3 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 A.6(i) is a direct result of Lemma A.1 if we note that for each $1\le l\le p$, $h_n^{-2}K'\left(\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right) - z\left(\mathbf{X}_{e,j},\boldsymbol{\beta}\right)\right)/h_n\right)\left(X_l - X_{l,j}\right)\cdot\left(\cdot_j\right)$ 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 A.6(ii), we note that

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

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 $\int K'(v)\,dv = K(v)|_{-\infty}^{\infty} = 0$, $\int vK'(v)\,dv = K(v)|_{-\infty}^{\infty} - \int K(v)\,dv = -1$, $\int v^sK'(v)\,dv = v^sK(v)|_{-\infty}^{\infty} - s\int v^{s-1}K(v)\,dv = 0$ for $s=2,3$, and $\left|\int v^4K'(v)\,dv\right| < \infty$. 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 A.6(ii).

Finally, Lemma A.6(iii) is a combination of Lemma A.6(i) and Lemma A.6(ii). □

Lemma A.7. *Suppose that p is fixed. If all the assumptions in Lemma A.3 hold with $\upsilon_G = 4$, $\upsilon_K = 3$, and $\upsilon_f = 4$, we have that*

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

where $\mathcal{X}_e^\phi$ is defined in (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

$$\left\| \frac{A'_{n,y}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta})} - \frac{\partial H_1(z(\mathbf{X}_e,\boldsymbol{\beta}),\mathbf{X}_e)/\partial z}{f_z(z(\mathbf{X}_e,\boldsymbol{\beta}))} \right\| = \left\| \frac{A'_{n,y}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta})} - \frac{A'_{y}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{1}(\mathbf{X}_e,\boldsymbol{\beta})} \right\|$$

$$\leq \left\| \frac{A'_{n,y}(\mathbf{X}_e,\boldsymbol{\beta}) - A'_{y}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta})} \right\| \tag{7}$$

$$+ \left\| \frac{A'_{y}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{1}(\mathbf{X}_e,\boldsymbol{\beta})} \frac{A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta}) - A_{1}(\mathbf{X}_e,\boldsymbol{\beta})}{A_{n,1}(\mathbf{X}_e,\boldsymbol{\beta})} \right\|. \tag{8}$$

Now for any $(\mathbf{X}_e,\boldsymbol{\beta}) \in \mathcal{X}_e^{\phi}\times\mathcal{B}$, $A_1(\mathbf{X}_e,\boldsymbol{\beta})$ is uniformly lower-bounded according to Lemma A.4, so $A_{n,1}^{-1}(\mathbf{X}_e,\boldsymbol{\beta}) = O_p(1)$ also uniformly holds. Moreover, $\left\|A'_y(\mathbf{X}_e,\boldsymbol{\beta})\right\|$ is upper bounded, so $\left\|A'_{n,y}(\mathbf{X}_e,\boldsymbol{\beta})\right\| = O_p(1)$ also uniformly holds. Then (7) is $O_p\left(h_n^{-2}\sqrt{\log\left(nh_n^{-1}\right)/n} + h_n^3\right)$ and (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

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

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 A.8. *Suppose that p is fixed. If all the assumptions in Lemma A.3 hold with $\upsilon_G = 4$, $\upsilon_K = 3$, and $\upsilon_f = 4$, then for any $\overline{\mathcal{B}} \subseteq \mathcal{B}$, we have that*

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

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

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

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 A.2(viii). Note that

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

for some $C > 0$ due to Assumption 4(i) and the choice of $\mathcal{X}_e^\phi$, so we have that

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

This proves the results. □

Now we prove Lemma 2 in the main text.

Proof of Lemma 2. Note that

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\phi \frac{\partial\widehat{G}(z(\mathbf{X}_{e,i},\boldsymbol{\beta})|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}-\Lambda_\phi(\boldsymbol{\beta})\right\|\\
\leq&\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\phi\left(\frac{\partial\widehat{G}(z(\mathbf{X}_{e,i},\boldsymbol{\beta})|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}-\int W(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta})(\mathbf{X}_i-\mathbf{X})\,d\mathbf{X}\right)\right\| \qquad (9)\\
&+\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\phi\left(\int W(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta})(\mathbf{X}_i-\mathbf{X})\,d\mathbf{X}\right)-\Lambda_\phi(\boldsymbol{\beta})\right\|. \qquad (10)
\end{aligned}$$

Obviously, (9) is of order $O_p\left(h_n^{-2}\sqrt{\log(nh_n^{-1})/n}+h_n^3+\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\|\Delta\boldsymbol{\beta}\|\right)$ according to Lemma A.8. Using Lemma A.1, we can show that (10) is $O_p\left(\sqrt{(\log n)/n}\right)$ by noting that each element of $\int W(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta})(\mathbf{X}_i-\mathbf{X})\,d\boldsymbol{X}$ 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 A.2(x). This finishes the proof of Lemma 2. □

Now we prove Lemma 3 in the main text.

Proof of Lemma 3. We first show that

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

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

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

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

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

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

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

where the RHS is equivalent to

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

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

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

So

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

where the bound does not depend on i. So

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

On the other side, we have that we have that

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

Together we have that

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

So

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

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,

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left[1-\frac{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)}{f_z^\star\left(z_i^\star\right)}\right]X_i^{\phi 2}\right]$$

$$= \frac{1}{n}\mathbb{E}\left\{G\left(z_i^\star\right)\left(1-G\left(z_i^\star\right)\right)\left[1-\frac{\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right)}{f_z^\star\left(z_i^\star\right)}\right]^2 X_i^{\phi 2}\right\}$$

$$\leq \frac{C}{n}\mathbb{E}\left\{\left(\frac{1}{n}\sum_{j=1}^{n}K_{h_n}\left(z_j^\star - z_i^\star\right) - f_z^\star\left(z_i^\star\right)\right)^2 X_i^{\phi 2}\right\}$$

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

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

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

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

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

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

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

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

$$=-\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\left(\mathbf{X}_i^\phi-\mathbb{E}\left(\mathbf{X}_i^\phi\middle|\,z_i^\star\right)\right)+o_p\left(1\right).$$

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 A.9. *Suppose that Assumption 1, Assumption 2(i) and (ii), and Assumption 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 A.9. This is a direct result of Lemma A.1 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 A.10. *Suppose that Assumption 1, Assumption 2(i) and (ii), and Assumption 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 A.10. 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 $\min\left\{\underline{\lambda}_A\left\|B\right\|,\underline{\lambda}_B\left\|A\right\|\right\}\leq$

$\|AB\| \leq \max\left\{\overline{\lambda}_A \|B\|, \overline{\lambda}_B \|A\|\right\}$, so we have that

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

□

Lemma A.11. *Suppose that Assumption 1, Assumption 2(i) and (ii), and Assumption 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

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

Proof of Lemma A.11. Note that

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

For the first term, we have that

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

For the second term, we note that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}}\sup_{\boldsymbol{\beta}\in\mathcal{B}} \left\|\mathfrak{X}_q(z,\boldsymbol{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_q^{-1}(\boldsymbol{\beta})\boldsymbol{r}_q(z)\mathbf{X}_i\right)\right\| \\
&\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}}\sup_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}}\sup_{\mathbf{X}_e\in\mathcal{X}_e} \left\|\mathfrak{X}_q\left(X_0 + \mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}},\boldsymbol{\beta}\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{r}_q^{\mathrm{T}}\left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\Gamma_q^{-1}(\boldsymbol{\beta})\boldsymbol{r}_q\left(X_0 + \mathbf{X}^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}\right)\mathbf{X}_i\right)\right\|,
\end{aligned}$$

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

$$\left| \boldsymbol{r}_q^{\mathrm{T}} \left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right) \Gamma_q^{-1} \left(\boldsymbol{\beta} \right) \boldsymbol{r}_q \left(X_0 + \mathbf{X}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} \right) X_{i,j} \right| \leq C q D_{q,0}^2,$$

and

$$\left\| \frac{\partial \boldsymbol{r}_q^{\mathrm{T}} \left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right) \Gamma_q^{-1} \left(\boldsymbol{\beta} \right) \boldsymbol{r}_q \left(X_0 + \mathbf{X}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} \right) X_{i,j}}{\partial \mathbf{X}_e} \right\| \leq C \sqrt{p} q D_{q,0} D_{q,1},$$

$$\left\| \frac{\partial \boldsymbol{r}_q^{\mathrm{T}} \left(X_{0,i} + \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right) \Gamma_q^{-1} \left(\boldsymbol{\beta} \right) \boldsymbol{r}_q \left(X_0 + \mathbf{X}^{\mathrm{T}} \widetilde{\boldsymbol{\beta}} \right) X_{i,j}}{\partial \widetilde{\boldsymbol{\beta}}} \right\| \leq C \sqrt{p} q D_{q,0} D_{q,1},$$

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

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

Lemma A.12. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold with $\upsilon_G \geq 1$, and that $\chi_{1,n} \to 0$ as $n \to \infty$, then we have that*

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

Proof of Lemma A.12. 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

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

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 A.13. *Suppose that Assumption 1, Assumption 2(i)-(iii), and Assumption 6 hold with $\upsilon_G \geq 1$, and that $\chi_{1,n} \to 0$ as $n \to \infty$, then we have that*

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

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

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 A.13. For the first result, we note that

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

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

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

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 4 in the main text.

Proof of Lemma 4. We note that

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

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 $\sup_{k\geq 1}\|\mathfrak{R}_{n,k}\|$, which is done based on Lemma A.11, Lemma A.12, and Lemma A.13. □

Now we prove Lemma 5 and Lemma 6 in the main text.

Proof of Lemma 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 A.11, 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,

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

For the second term, we have that

$$\sup_{0\le t\le 1,\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}G'\left(z_i^\star+t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)\left(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}-\mathfrak{X}_{q,n}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}_i^{\mathrm{T}}\right)-\Psi_q\left(t,\boldsymbol{\beta}\right)\right\|$$
$$=O_p\left(\sqrt{p^3q^2D_{q,0}^4\log\left(pqD_{q,0}D_{q,1}n\right)/n}\right)=O_p\left(p\chi_{1,n}\right),$$

due to the fact that

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

and

$$\begin{aligned}&\left|G'\left(z_i^\star+t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_1\right)\left(X_{i,s}X_{i,t}-\left(\mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_1\right),\boldsymbol{\beta}_1\right)\right)_s X_{i,t}\right)\right.\\&\left.-G'\left(z_i^\star+t\mathbf{X}_i^{\mathrm{T}}\Delta\boldsymbol{\beta}_2\right)\left(X_{i,s}X_{i,t}-\left(\mathfrak{X}_q\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}_2\right),\boldsymbol{\beta}_2\right)\right)_s X_{i,t}\right)\right|\\&\le C\sqrt{p}q^2D_{q,0}^3D_{q,1}\left\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\right\|.\end{aligned}$$

Finally,

$$\begin{aligned}&\sup_{0\le t\le 1,\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\Psi_q\left(t,\boldsymbol{\beta}\right)-\Psi_q^\star\right\|\\&\le\sup_{0\le t\le 1,\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\mathbb{E}\left[G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^\star\right)+t\mathbf{X}^{\mathrm{T}}\Delta\boldsymbol{\beta}\right)-G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^\star\right)\right)\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}-\mathfrak{X}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)\mathbf{X}^{\mathrm{T}}\right)\right]\right\|\\&+\sup_{\boldsymbol{\beta}\in\mathcal{B}_n}\left\|\mathbb{E}\left[G'\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^\star\right)\right)\left(\mathfrak{X}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right)-\mathfrak{X}_q\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}^\star\right),\boldsymbol{\beta}^\star\right)\mathbf{X}^{\mathrm{T}}\right)\right]\right\|.\end{aligned}$$

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

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

So

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

Combine the above results, we have that

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

□

Proof of Lemma 6. According to Theorem 7, we have that $\sup_{k\ge k_{1,n}^{SBGD}+1} \|\Delta\boldsymbol{\beta}_k\| = O_p(\chi_{2,n})$. To prove the lemma, we first show that

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

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

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

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

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

which leads to the result if we further note that

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

Next we show that

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

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

The first term is obviously bounded in probability by

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

The second term is bounded by

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

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

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

$\square$

B Proofs of Theorems

Proof of Theorem 1

Proof. We first prove Theorem 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 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 A.1, 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(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 $\tau > 0$,

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

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(i).

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

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

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(ii)

Now we prove Theorem 1(iii). We first note that for any square matrices A, B, and C, there hold $\|AB\| \le \overline{\sigma}(A)\|B\|$ and $\|ABC\| \le \overline{\sigma}(A)\|BC\| \le \overline{\sigma}(A)\overline{\sigma}(B)\|C\|$. So

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

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

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

On the other side, we have that

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

Together, we have that

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

Note that $\overline{\lambda}(M^{-1}(\boldsymbol{\beta}_e^\star)) \le \underline{\lambda}_e^{-1}$, $\overline{\lambda}\left(\mathbb{E}\left[G_i^\star(1-G_i^\star)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right) \le \frac{1}{4}\overline{\lambda}\left(\mathbb{E}\left[\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right]\right) \le C$, $\overline{\lambda}\left(M_n^{-1}\left(\widehat{\boldsymbol{\beta}}_e\right)\right) \le$

$2\underline{\lambda}_e^{-1}$ with probability going to 1, and $\overline{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i\left(1-\widehat{G}_i\right)\mathbf{X}_{e,i}\mathbf{X}_{e,i}^{\mathrm{T}}\right)\le C$ with probability going to 1, we have that

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

which validates the result.

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

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

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

Proof of Theorem 2

Proof. We first show Theorem 2(i). Note that from the proof in Theorem 1, we know that with probability going to 1, we have that

$$\begin{aligned}\left\|\Delta\boldsymbol{\beta}_{e,k+1}\right\| &\le \sup_{\boldsymbol{\beta}_e\in\mathcal{B}}\overline{\lambda}\left(I_{p+1}-\delta_k M_n\left(\boldsymbol{\beta}_e\right)\right)\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|+\delta_k\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\| \\ &\le \left(1-\underline{\lambda}_e\delta_k/2\right)\left\|\Delta\boldsymbol{\beta}_{e,k}\right\|+\delta_k\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\|\le\cdots \\ &\le \left(\prod_{j=1}^{k}\left(1-\underline{\lambda}_e\delta_j/2\right)\right)\left\|\Delta\boldsymbol{\beta}_{e,1}\right\|+\left\{\sum_{j=0}^{k-1}\delta_{k-j}\left(\prod_{l=0}^{j-1}\left(1-\underline{\lambda}_e\delta_{k-l}/2\right)\right)\right\}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{X}_{e,i}\right\|,\end{aligned} \tag{11}$$

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 (11), since $e^x\ge 1+x$ for all x, we have $1-\underline{\lambda}_e\delta_j/2\le\exp\left(-\underline{\lambda}_e\delta_j/2\right)$ for all j. Define $S_0=0$ and $S_j=\sum_{l=1}^{j}\delta_l$ for $j\ge 1$, we have that

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

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

$$\sum_{j=0}^{k-1} \delta_{k-j} \left(\prod_{l=0}^{j-1} (1 - \underline{\lambda}_e \delta_{k-l}/2) \right) \leq \sum_{j=0}^{k-1} \delta_{k-j} \exp\left(-\frac{\underline{\lambda}_e}{2} \sum_{l=0}^{j-1} \delta_{k-l} \right)$$

$$\leq C \sum_{j=0}^{k-1} \delta_{k-j+1} \exp\left(-\frac{\underline{\lambda}_e (S_k - S_{k-j})}{2} \right)$$

$$= C \exp\left(-\frac{\underline{\lambda}_e S_k}{2} \right) \sum_{j=0}^{k-1} (S_{k-j+1} - S_{k-j}) \exp\left(\frac{\underline{\lambda}_e S_{k-j}}{2} \right)$$

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

Then we have that

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

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

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

and

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

so $\left\| \Delta \boldsymbol{\beta}_{e,k+1} \right\| = O_p\left(\sqrt{p (\log p)/n} \right)$. This validates Theorem 2(i).

For Theorem 2(ii), we know that for $k \geq \widetilde{k}_{1,n}^{BGD} + 1$, $\left\| \Delta \boldsymbol{\beta}_{e,k} \right\| = O_p\left(\sqrt{p (\log p)/n} \right)$ holds, so we have that

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

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

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, we can easily show that

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

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

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

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 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

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

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 2(ii).

The proof of Theorem 2(iii) and Theorem 2(iv) is the same as that in the proof of Theorem 1, so is left out. □

Proof of Theorem 3

Proof. Define

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

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

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 $0 \le t \le 1$, so

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

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

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

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

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

Then according to Assumption 5, we have that

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

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

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

So

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

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, 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 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

$$
\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\Lambda_{\phi}\left(\boldsymbol{\beta}\right) - \Lambda\left(\boldsymbol{\beta}\right)\right\| \leq \delta\underline{\lambda}_{\Lambda}/8 \tag{12}
$$

holds under the choice of ϕ.

Based on (12), the following proof is similar to the proof of Theorem 3. Define

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

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

We have that

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

so

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

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 (12) holds, so we have that

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

According to the proof of Theorem 3, there holds $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\overline{\sigma}\left(I_p - \delta\mathit{\Lambda}\left(\boldsymbol{\beta}\right)\right) \leq 1 - \delta\underline{\lambda}_{\mathit{\Lambda}}/4$ under the choice of δ, so we have that

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

Then based on the proof of Theorem 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. $\square$

Proof of Theorem 5

Proof. Note that under the choice of δ and ϕ , $\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1}\|\boldsymbol{\beta}_k-\boldsymbol{\beta}^{\star}\|=O_p\left(\psi\left(n,p,h_n\right)\right)$ according to Theorem 4. According to (14), we have that

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

According to Lemma 2, we have that

$$\begin{aligned}\sup_{k\geq\widetilde{k}_{1,n}^{KBGD},t\in[0,1]}&\left\|\left\{I_p-\frac{\delta}{n}\sum_{i=1}^{n}\left[\mathbf{X}_i^{\phi}\frac{\partial\widehat{G}\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)|\,\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k}\right]\right\}\right.\\&\left.-\left\{I_p-\delta\Lambda_{\phi}\left(\boldsymbol{\beta}^{\star}+t\Delta\boldsymbol{\beta}_k\right)\right\}\right\|=\delta O_p\left(h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n}+h_n^3\right),\end{aligned}\tag{13}$$

due to the fact that

$$\sup_{k\geq\widetilde{k}_{1,n}^{KBGD}+1}\|\Delta\boldsymbol{\beta}_k\|=O_p\left(\psi_1\left(n,p,h_n\right)\right)=o_p\left(h_n^{-2}\sqrt{\left(\log\left(nh_n^{-1}\right)\right)/n}+h_n^3\right),$$

when p is fixed and $h_n\to 0$.

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 (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 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 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

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

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 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 $I_i^\phi\cdot\int_{\mathcal{X}}V\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)d\mathbf{X}$ 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 $f_z\left(z\left(\mathbf{X}_{e,i},\boldsymbol{\beta}\right)\middle|\,\boldsymbol{\beta}\right)>0$, so $\int_{\mathcal{X}}\left\|\partial W\left(\mathbf{X}_{e,i},\mathbf{X}_e,\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta}\right\|d\mathbf{X}$ is bounded according to Lemma A.2(x). This implies that

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

Then

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

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

$$\left\|\left(I_p - \delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^k \Delta\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD}+k_{2,n}^{KBGD}+1}\right\| \leq \left(1-\delta\underline{\lambda}_\Lambda/8\right)^k \left\|\Delta\boldsymbol{\beta}_{\widetilde{k}_{1,n}^{KBGD}+k_{2,n}^{KBGD}+1}\right\|,$$

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

$$\left\|\Lambda_\phi^{-1}\left(\boldsymbol{\beta}^\star\right)\boldsymbol{\xi}_n^\phi - \delta\sum_{j=1}^{k}\left(I_p - \delta\Lambda_\phi\left(\boldsymbol{\beta}^\star\right)\right)^{k-j}\boldsymbol{\xi}_n^\phi\right\| \leq 8\lambda_\Lambda^{-1}\left(1-\delta\underline{\lambda}_\Lambda/8\right)^{k+1}\left\|\boldsymbol{\xi}_n^\phi\right\|.$$

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

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

According to Lemma 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 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 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. In particular, let $\phi_n \downarrow 0$ and $\mathcal{X}_{e,n}$ be as defined as in the proof of Lemma 1. Then we have that $f_z^\star(z_i^\star) \geq C\phi_n^p$ as long as $\mathbf{X}_{e,i} \in \mathcal{X}_{e,n}$. Denote $G_i^\star = G(z_i^\star)$, we have

$$\begin{aligned}
\left\|\widehat{\Sigma}_{\boldsymbol{\xi}}^\phi - \Sigma_{\boldsymbol{\xi}}^\phi\right\| \leq & \left\| \frac{1}{n}\sum_{i=1}^n \left(I_{n,i} \cdot \widehat{G}_i \left(1-\widehat{G}_i\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| \widehat{z}_i\right)\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| \widehat{z}_i\right)\right)^{\mathrm{T}}\right) \right. \\
& \left. -\mathbb{E}\left(I_{n,i} \cdot G_i^\star \left(1-G_i^\star\right) \left(\mathbf{X}_i^\phi - \mathbb{E}\left(\mathbf{X}_i^\phi \middle| z_i^\star\right)\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| z_i^\star\right)\right)^{\mathrm{T}}\right) \right\| \qquad (14) \\
& + \left\| \frac{1}{n}\sum_{i=1}^n \left(\left(1-I_{n,i}\right) \cdot \widehat{G}_i \left(1-\widehat{G}_i\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| \widehat{z}_i\right)\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| \widehat{z}_i\right)\right)^{\mathrm{T}}\right) \right. \\
& \left. -\mathbb{E}\left(\left(1-I_{n,i}\right) \cdot G_i^\star \left(1-G_i^\star\right) \left(\mathbf{X}_i^\phi - \mathbb{E}\left(\mathbf{X}_i^\phi \middle| z_i^\star\right)\right) \left(\mathbf{X}_i^\phi - \widehat{\mathbb{E}}\left(\mathbf{X}_i^\phi \middle| z_i^\star\right)\right)^{\mathrm{T}}\right) \right\|. \qquad (15)
\end{aligned}$$

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 (15) is $O_p(\phi_n)$. Now we look at (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 A.7, we have that

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

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 $\left\|L\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\boldsymbol{\beta}\right) f_z^{-1}\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right)\right)\partial H_2\left(z\left(\mathbf{X}_e,\boldsymbol{\beta}\right),\mathbf{X}_e\right)/\partial z\right\|$ 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

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

So we have that (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 7

Proof. The proof is similar to that of Theorem 3. Note that

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

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

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

so

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

and

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

Then we have that

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

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

Proof of Theorem 8

Proof. We first prove Theorem 8 (i). Note that

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

Define

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

According to Lemma 5, we have that

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

According to Lemma 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 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

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

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

and

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

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

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

The following results hold trivially. $\square$

Proof of Theorem 9

Proof. Note that under all the conditions imposed in Theorem 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 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 A.11, 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

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

Note that the above is bounded by

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

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 $\upsilon_G\ge 2$, we have that

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

So

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

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(\rho)-\sigma_S^2(\rho)\right| \\
&\leq\|\rho\|^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\|\rho\|^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\| \\
&+\|\rho\|^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\| \\
&+\|\rho\|^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

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

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

References

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.

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.

Herman J Bierens. Consistency and asymptotic normality of sieve ml estimators under low-level conditions. *Econometric Theory*, 30(5):1021–1076, 2014.

M.D. Cattaneo, M. Jansson, and W.K. Newey. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34:277–301, 2018.

Christopher Cavanagh and Robert P Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351–382, 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.

Yanqin Fan, Fang Han, Wei Li, and Xiao-Hua Zhou. On rank estimators in increasing dimensions. *Journal of Econometrics*, 214(2):379–412, 2020.

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.

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.

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.

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

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.

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.

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.

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.

Haiyong Wang and Shuhuang Xiang. On the convergence rates of legendre approximation. *Mathematics of computation*, 81(278):861–877, 2012.

Qingsong Yao. *Kernel-Based Learning of Monotone Index Models with Large Dimensionality*. PhD thesis, Boston College, 2023.